Soft Cardinality: A Parameterized Similarity Function for Text Comparison

Sergio Jimenez
Universidad Nacional
de Colombia, Bogota,
Ciudad Universitaria
edificio 453, oficina 220
sgjimenezv@unal.edu.co

Claudia Becerra
Universidad Nacional
de Colombia, Bogota
cjbecerrac@unal.edu.co

Alexander Gelbukh
CIC-IPN
Av. Juan Dios Bátiz,
Av. Mendizábal, Col.
Nueva Industrial Vallejo,
CP 07738, DF, México
gelbukh@gelbukh.com

Abstract

We present an approach for the construction of text similarity functions using a parameterized resemblance coefficient in combination with a softened cardinality function called soft cardinality. Our approach provides a consistent and recursive model, varying levels of granularity from sentences to characters. Therefore, our model was used to compare sentences divided into words, and in turn, words divided into q-grams of characters. Experimentally, we observed that a performance correlation function in a space defined by all parameters was relatively smooth and had a single maximum achievable by "hill climbing." Our approach used only surface text information, a stop-word remover, and a stemmer to tackle the semantic text similarity task 6 at SEMEVAL 2012. The proposed method ranked 3rd (average), 5th (normalized correlation), and 15th (aggregated correlation) among 89 systems submitted by 31 teams.

1 Introduction

Similarity is the intrinsic ability of humans and some animals to balance commonalities and differences when comparing objects that are not identical. Although there is no direct evidence of how this process works in living organisms, some models have been proposed from the cognitive perspective (Sjöberg, 1972; Tversky, 1977; Navarro and Lee, 2004). On the other hand, several similarity models have been proposed in mathematics, statistics, and computer science among other fields. Particularly in AI, similarity measures play an important role in the construction of intelligent systems that are required to exhibit behavior similar to humans. For instance, in the field of natural language processing, text similarity functions provide estimates of the human similarity judgments related to language. In this paper, we combine elements from the perspective of cognitive psychology and computer science to propose a model for building similarity functions suitable for the task of semantic text similarity.

We identify four main families of text similarity functions: i) resemblance coefficients based on sets (e.g. Jaccard's (1901) and Dice's (1945) coefficients) ii) functions in metric spaces (e.g. cosine *tf-idf* similarity (Salton et al., 1975)); iii) the edit distance family of measures (e.g. Levenstein (1966) distance, LCS (Hirschberg, 1977)); and iv) hybrid approaches ((Monge and Elkan, 1996; Cohen et al., 2003; Corley and Mihalcea, 2005; Jimenez et al., 2010)). All of these measures use a subdivision of the texts in different granularity levels, such as q-grams of words, words, q-grams of characters, syllables, and characters. Among hybrid approaches, Monge-Elkan's measure and soft cardinality methods are recursive and can be used to build similarity functions at any arbitrary range of granularity. For instance, it is possible to construct a similarity function to compare sentences based on a function that compares words, which in turn can be constructed based on a function that compares bigrams of characters. Furthermore, hybrid approaches can integrate similarity functions that are not based on the representation of the surface of text, such as semantic relatedness measures (Pedersen et al., 2004).

Text similarity measures can be static or adaptive whether they are binary functions using only surface information of the two texts, or are functions that suit to a wider set of texts. For instance, measures using *tf-idf* weights adapt their results to the set of texts in which those weights were obtained. Other approaches learn parameters of the similarity function from a set of texts to optimize a particular task. For instance, Ristad and Yianilos (1998) and Bikenko and Mooney (2003) learned the costs of edit operations for all characters for an edit-distance function in a name-matching task. Other machine-learning approaches have also been proposed to build adaptive measures in name-matching (Bilenko and

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 449–453,
Montréal, Canada, June 7-8, 2012.

Mooney, 2003) and textual-entailment tasks.

However, those machine-learning-based methods for adaptive similarity suffer from sparseness and the "curse of dimensionality". For example, the method of Ristad and Yianilos learns $n^2 + 2n$ parameters, where n is the size of the character set. Similarly, dimensionality in the method of Bilenko and Mooney is the size of the data set vocabulary. This issue is addressed primarily through machine-learning algorithms, which reduce the dimensionality of the problem regularizating to achieve enough generalization to get an acceptable performance difference between training and test data. Although machine-learning solutions have proven effective for many applications, the principle of Occam's razor suggests that it should be preferable to have a model that explains the data with a smaller number of significant parameters. In this paper, we seek a simpler adaptive similarity model with few meaningful parameters.

Our proposed similarity model starts with a cardinality-based resemblance coefficient (i.e. Dice's coefficient $2|A \cap B|/|A|+|B|$) and generalizes it to model the effect of asymmetric selection of the referent. This effect is a human factor discovered by Tversky (1977) that affects judgments of similarity, i.e. humans tends to select the more prominent stimulus as the referent and the less salient stimulus as the object. Some of Tversky's examples are "the son resembles the father" rather than "the father resembles the son", "an ellipse is like a circle" not "a circle is like an ellipse", and "North Korea is like Red China" rather than "Red China is like North Korea". Generally speaking, "the variant is more similar to the prototype than vice versa". In the previous example, stimulus salience is associated with the prominence of the country; for text comparison we associate word salience with *tf-idf* weights. At the text level, we associate salience with a combination of word-salience, inter-word similarity, and text length provided by soft cardinality. Experimentally, we observed that this effect also occurs when comparing texts, but not necessarily in the same direction suggested by Tversky. We used this effect to improve the performance of our similarity model. In addition, we proposed a parameter that biases the function to generate greater or lower similarity scores.

Finally, in our model we used a soft cardinality function (Jimenez et al., 2010) instead of the classical set cardinality. Just as classical cardinality counts the number of elements which are not identical in a set, soft cardinality uses an auxiliary inter-element similarity function to make a soft count. For instance, the soft cardinality of a set with two very similar (but not identical) elements should be a real number closer to 1.0 instead of 2.0.

The rest of the paper is organized as follows. In Section 2 we briefly present soft cardinality. In Section 3 the proposed parameterized similarity model is presented. In Section 4 experimental validation is provided using 8 data sets annotated with human similarity judgments from the "Semantic-Text-Similarity" task at SEMEVAL-2012. Finally, a brief discussion is provided in Section 5 and conclusions are presented in Section 6.

2 Soft Cardinality

Let $A = \{a_1, a_2, \ldots, a_{|A|}\}$ and $B = \{b_1, b_2, \ldots, b_{|B|}\}$ be two sets being compared. When each element of a_i or b_j has an associated weight w_{a_i} or w_{b_j} the problem of comparing those sets becomes a weighted similarity problem. This means that such model has to take into account not only the commonalities and diferences, but also their weights. Also, if an $(|A \cup B|) \times (|A \cup B|)$ similarity matrix $\mathbf{S}$ is available, the problem becomes a weighted soft similarity problem because the commonality between A and B has to be computed not only with identical elements, but also with elements with a degree of similarity. The values of $\mathbf{S}$ can be obtained from an auxiliary similarity function $sim(a, b)$ that satisfies at least non-negativity ($\forall a, b,\ sim(a, b) \geq 0$) and reflexivity ($\forall a,\ sim(a, a) = 1$). Other postulates such as symmetry ($\forall a, b,\ sim(a, b) = sim(b, a)$) and triangle inequality[1] ($\forall a, b, c,\ sim(a, c) \geq sim(a, b) + sim(b, c) - 1$) are not strictly necessary.

Jimenez et al. (2010) proposed a set-based weighted soft-similarity model using resemblance coefficients and the soft cardinality function instead of classical set cardinality. The idea of calculating the soft cardinality is to treat elements a_i in set the A as sets themselves and to treat inter-element similarities as the intersections between the elements $sim(a_i, a_j) = |a_i \cap a_j|$. Therefore, the soft cardinality of set A becomes $|A|' = \left|\bigcup_{i=1}^{|A|} a_i\right|$. Since it is not feasible to calculate this union, they proposed the following weighted approximation using $|a_i| = w_{a_i}$:

$$|A|'_{sim} \simeq \sum_{i}^{|A|} w_{a_i} \left(\sum_{j}^{|A|} sim(a_i, a_j)^p \right)^{-1} \quad (1)$$

Parameter $p \geq 0$ in eq.1 controls the "softeness" of the cardinality, taking $p = 1$ its no-effect value and leaving element similarities unchanged for the calculation of soft cardinality. When p is large, all $sim(*, *)$ results lower than 1 are transformed into a number approaching 0. As a result, the soft cardinality behaves like the classical cardinality, returning the addition of all the weights of the elements, i.e $|A|'_{sim} \simeq \sum_i^{|A|} w_{a_i}$. When p is close to 0, all $sim(*.*)$ results are transformed approaching

[1] triangle inequality postulate for similarity is derived from its counterpart for dissimilarity (distance) $distance(a, b) = 1 - sim(a, b)$.

into a number approaching 1, making the soft cardinality returns the average of the weights of the elements, i.e. $|A|'_{sim} \simeq \frac{1}{|A|}\sum_{i}^{|A|} w_{a_i}$. Jimenez et al. used $p = 2$ and *idf* weights in the same name-matching task proposed by Cohen et al. (Cohen et al., 2003).

3 A Parameterized Similarity Model

As we mentioned above, Tvesky proposed that humans tends to select more salient stimulus as *referent* and less salient stimulus as *object* when comparing two objects A and B. Based on the idea of Tvesrky, the similarity between two objects can be measured as the ratio between *the salience of commonalities* and *the salience of the less salient object*. Drawing an analogy between objects as sets and salience as the cardinality of a set, the salience of commonalities is $|A \cap B|$, and the salience of the less salient object is $\min(|A|, |B|)$. This ratio is known as the overlap coefficient $Overlap(A, B) = \frac{|A \cap B|}{\min(|A|,|B|)}$. However, whether $|A| < |B|$ or whether $|A| \ll |B|$, the similarity obtained by $Overlap(A, B)$ is the same. Hence, we propose to model the selecction of the referent using a parameter α that makes a weighted average between $\min(|A|, |B|)$ and $\max(|A|, |B|)$, controling the degree to which the asymmetric referent-selection effect is considered in the similarity measure.

$$SIM(A, B) = \frac{|A \cap B| + bias}{\alpha \max(|A|, |B|) + (1 - \alpha) \min(|A|, |B|)} \quad (2)$$

The parameter α controls the degree to which the asymmetric referent-selection effect is considered in the similarity measure. Its no-effect value is $\alpha = 0.5$, so the eq.2 becomes the Dice coefficient. Moreover, when $\alpha = 0$ the eq.2 becomes the overlap coefficient, otherwise when $\alpha = 1$ the opposite effect is modeled.

In addition, we introduced a $bias$ parameter in eq. 2 that increases the commonalities of each object pair by the same amount, and so it measures the degree to which all of the objects have commonalities among each other. Clearly, the non-effect value for the $bias$ parameter is 0.

Besides, the $bias$ parameter has the effect of biasing $SIM(A, B)$ by considering any pair $\langle A, B\rangle$ more similar if $bias > 0$ and their cardinalities are small. Conversely, the similarity between pairs with large cardinalities is promoted if $bias < 0$. However, as higher values of $bias$ may result in similarity scores outside the interval $[0, 1]$, additional post-procesing to limit the similarities in this interval may be required.

The proposed parameterized text similarity measure is constructed by combining the proposed resemblance coefficient in eq.2 and the soft cardinality in eq.1. The resulting measure has three parameters: α, $bias$, and p. Weights w_{a_i} can be *idf* weights. This measure takes two texts represented as sets of words and returns their similarity. The auxiliary similarity function $sim(a, b)$ necessary for calculating the soft cardinality is another parameter of the model. This auxiliary function is any function that can compare two words and return a similarity score in $[0, 1]$.

α	Asymetric referent selection at text level
$bias$	Bias parameter at text level
p	Soft cardinality exponent at word level
w_{a_i}	Element weights at word level
q_1, q_2	q_1-grams or $[q_1 : q_2]$spectra word division
α_{sim}	Asymetric referent selection at q-gram level
$bias_{sim}$	Bias parameter q-gram level

Table 1: Parameters of the proposed similarity model

To build this $sim(a, b)$ function, we chose to reuse the eq.2 but representing words as sets of q-grams or ranges of q-grams of different sizes, i.e. $[q_1 : q_2]$ spectra. Q-grams are consecutive overlapped substrings of size q. For instance, the word "*saturday*" divided into trigrams is {◁*sa*, s*at*, *atu*, *tur*, *urd*, *rda*, *day*, *ay*▷}. The character '▷' is a padding character added to differenciate q-grams at the begining or end of the string. A $[2 : 4]$spectra is the combined representation of a word using –in this example– bigrams, trigrams and quadgrams (Jimenez and Gelbukh, 2011). The cardinality function for $sim()$ was the classical set cardinality. Clearly, the soft cardinality could be used again if an auxiliary similarity function for character comparison and a q-gram weighting mechanism are provided to allow another level of recursion. Therefore, the parameters of $sim(a, b)$ are: α_{sim}, $bias_{sim}$. Finally, the entire set of parameters of the proposed similarity model is shown in Table 1.

4 Experimental Setup and Results

The aim of these experiments is to observe the behavior of the parameters of our similarity model and verify if the hypothesis that motivated these parameters can be confirmed experimentally. The experimental data are 8 data sets (3 for training and 5 for test) proposed in the "Semantic Text Similarity" task at SEMEVAL-2012. Each data set consist of a set of pairs of text annotated with human-similarity judgments on a scale of 0 to 5. Each similarity judgment is the average of the judgments provided by 5 human judges. For a comprehensible description of the task see(Agirre et al., 2012).

For the experiments, all data sets were pre-processed by converting to lowercase characters, English stopwords removal and stemming using Porter stemmer (Porter, 1980). The performance measure used for all experiments was the Pearson correlation r.

4.1 Model Parameters

In order to make an initial exploration of the parameters in Table 1, we set $q_1 = 2$ (i.e. bigrams) and used $w_{a_i} = idf(a_i)$. For other parameters, we started with all the non-effect values, i.e. $\alpha = 0.5$, $bias = 0$, $p = 1$, $\alpha_{sim} = 0.5$ and $bias_{sim} = 0$. Plots in Figure 1 show the Pearson correlation measured in each of the data sets. For each graph, the non-effect configuration was used and each parameter varies in the range indicated in each horizontal axis. For best viewing, the non-effect values on each graph are represented by a vertical line.

In this exploration of the parameters it was noted that each parameter defines a function for the performance measure that is smooth and with an unique global maximum. Therefore, we assumed that the join performance function in the space defined by the 5 parameters also had the same properties. The parameters for each data set shown in Table 2 were found using a simple hill-climbing algorithm. Different q-gram and spectra configurations were tested manually.

5 Discussion

It is possible to observe from the results in Figure 1 and Table 2 that the behavior of the parameters is similar in pairs of data sets that have training and test parts. This behavior is evident in both MSRvid and MSRpar data sets, but it is less evident in SMTeuroparl. Furthermore, the optimal parameters for training data sets MSRvid and MSRpar were similar to those of their test data sets. In conclusion, the proposed set of parameters provides a set of features that characterize a data set for the text similarity task.

Regarding the effect of asymmetry in referent selecction proposed by Tvesrky, it was observed that –at text level– the MSRvid data sets were the only ones that supported this hypothesis ($\alpha = 0.32, 0.42$). The remaining data sets showed the opposite effect ($\alpha > 0.5$). That is, annotators chose the most salient document (the longer) as the referent when a pair of texts is being compared.

The Table 2 also shows that the optimal parameters for all data sets were different from the no-effect values combination. This result can also be seen in Figure 1, where curves crossed the vertical line of no-effect value –in most of the cases– in values different to the optimum. Clearly, the proposed set of parameters is useful for adjusting the similarity function for a particular data set and task.

6 Conclusions

We have proposed a new parameterized similarity function for text comparison and a method for finding the optimal values of the parameter set when training data is available. In addition, the parameter α, which was motivated by the similarity model of Tversky, proved effective in obtaining better performance, but we could not confirm the Tvesky's hypothesis that humans tends to select the object (text) with less stimulus salience (text length) as the referent. This result might have occurred because either the stimulus salience is not properly represented by the length of the text, or Tversky's hypothesis cannot be extended to text comparison.

The proposed similarity function proved effective in the task of "Semantic Text Similarity" in SEMEVAL 2012. Our method obtained the third best average correlation on the 5 test data sets. This result is remarkable because our method only used data from the surface of the texts, a stop-word remover, and a stemmer, which can be even be considered as a baseline method.

Acknowledgments

This research was funded by the Systems and Industrial Engineering Department, the Office of Student Welfare of the National University of Colombia, Bogotá, and throught a grant from the Colombian Department for Science, Technology and Innovation Colciencias, proj. 110152128465. The second author recognizes the support from Mexican Government (SNI, COFAA-IPN, SIP 20113295, CONACYT 50206-H) and CONACYT–DST India (proj. "Answer Validation through Textual Entailment").

References

Eneko Agirre, Daniel Cer, Mona Diab, and Gonzalez-Agirre Aitor. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proc. of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).*, Montreal,Canada.

Mikhail Bilenko and Raymond J. Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, Washington, D.C. ACM.

William W Cohen, Pradeep Ravikumar, and Stephen E Fienberg. 2003. A comparison of string distance metrics for Name-Matching tasks. In *Proc. of the IJCAI2003 Workshop on Information Integration on the Web II Web03*.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Stroudsburg, PA.

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, pages 297–302.

Daniel S. Hirschberg. 1977. Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675.

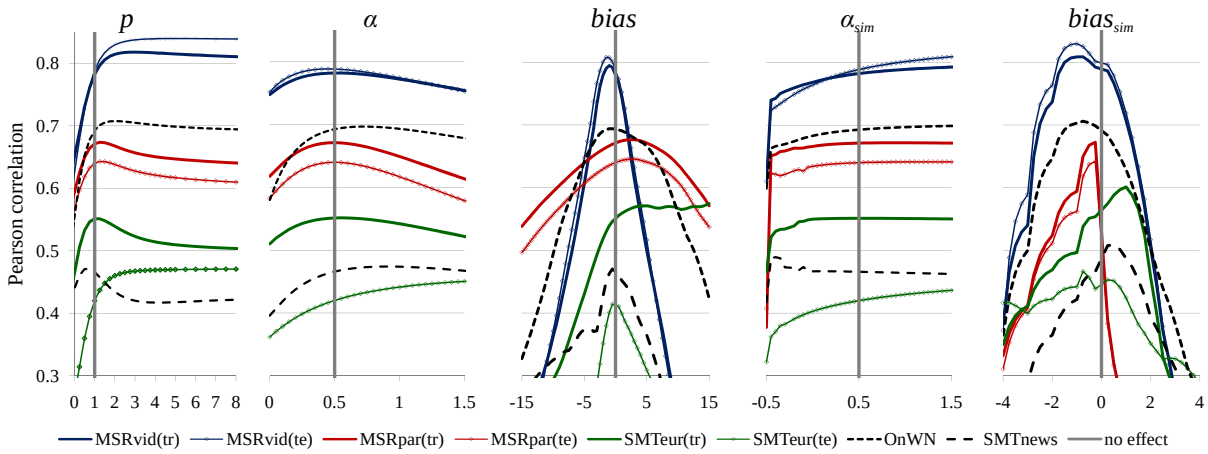

Figure 1: Exploring similarity model parameters around their no-effect values (tr=training, te=test)

	Parameters						correl.	Official Results	
Data set	$[q_1 : q_2]$	α	$bias$	p	α_{sim}	$bias_{sim}$	r	SoftCard	Best
MSRpar.training	[4]	0.62	1.14	0.77	-0.04	-0.38	0.6598	n/a	n/a
MSR.par.test	[4]	0.60	1.02	0.9	-0.02	-0.4	0.6335	0.6405[1]	0.7343
MSRvid.training	[1:4]	0.42	-0.80	2.28	0.18	0.08	0.8323	n/a	n/a
MSRvid.test	[1:4]	0.32	-0.80	1.88	1.08	0.08	0.8579	0.8562	0.8803
SMTeuroparl.training	[2:4]	0.74	-0.06	0.91	1.88	2.90	0.6193	n/a	n/a
SMTeuroparl.test	[2:4]	0.84	-0.16	0.71	1.78	3.00	0.5178	0.5152[2]	0.5666
OnWN.test	[2:5]	0.88	-0.62	1.36	-0.02	-0.70	0.7202	0.7109[1]	0.7273
SMTnews.test	[1:4]	0.88	0.88	1.57	0.80	3.21	0.5344	0.4833[1]	0.6085

[1]Result obtained using Jaro-Winkler (Winkler, 1990) measure as $sim(a, b)$ function between words.
[2]Result obtained using generalized Monge-Elkan measure $p = 4$, no stop-words removal and no term weights (Jimenez et al., 2009).

Table 2: Results with optimized parameters and official SEMEVAL 2012 results

Paul Jaccard. 1901. Etude comparative de la distribution florare dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, pages 547–579.

Sergio Jimenez and Alexander Gelbukh. 2011. SC spectra: a linear-time soft cardinality approximation for text comparison. In *Proc. of the 10th international conference on Artificial Intelligence*, MICAI'11, Puebla, Mexico.

Sergio Jimenez, Claudia Becerra, Alexander Gelbukh, and Fabio Gonzalez. 2009. Generalized Monge-Elkan method for approximate text string comparison. In *Computational Linguistics and Intelligent Text Processing*, volume 5449 of *LNCS*, pages 559–570.

Sergio Jimenez, Fabio Gonzalez, and Alexander Gelbukh. 2010. Text comparison using soft cardinality. In *String Processing and Information Retrieval*, volume 6393 of *LNCS*, pages 297–302.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Alvaro E. Monge and Charles Elkan. 1996. The field matching problem: Algorithms and applications. In *Proc. KDD-96*, pages 267–270, Portland, OR.

Daniel Navarro and Michael D. Lee. 2004. Common and distinctive features in stimulis representation: A modified version of the contrast model. *Psychonomic Bulletin & Review*, 11:961–974.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Proc. HLT-NAACL–Demonstration Papers*, Stroudsburg, PA.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 3(14):130–137.

Eric S. Ristad and Peter N. Yianilos. 1998. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Gerard Salton, A. Wong, and C.S. Yang. 1975. A vector space model for automatic indexing. *Com. ACM*, 18(11):613–620.

L. Sjöberg. 1972. A cognitive theory of similarity. *Göteborg Psychological Reports*.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proc. of the Section on Survey Research Methods*.