Similarity judgments: philosophical, psychological and mathematical investigations

Claude St-Jacques
Institute for Information Technology
National Research Council of Canada
Gatineau, QC, Canada
Claude.St-Jacques@nrc.gc.ca

Caroline Barrière
Institute for Information Technology
National Research Council of Canada
Gatineau, QC, Canada
Caroline.Barriere@nrc.gc.ca

Abstract

This study investigates similarity judgments from two angles. First, we look at models suggested in the psychology and philosophy literature which capture the essence of concept similarity evaluation for humans. Second, we analyze the properties of many metrics which simulate such evaluation capabilities. The first angle reveals that non-experts can judge similarity and that their judgments need not be based on predefined traits. We use such conclusions to inform us on how gold standards for word sense disambiguation tasks could be established. From the second angle, we conclude that more attention should be paid to metric properties before assigning them to perform a particular task.

1 Introduction

The task of word sense disambiguation has been at the heart of Natural Language Processing (NLP) for many years. Recent Senseval competitions (Mihalcea and Edmonds, 2004; Preiss and Yarowsky, 2001) have stimulated the development of algorithms to tackle different lexical disambiguation tasks. Such tasks require at their core a judgment of similarity as a word's multiple definitions and its contexts of occurrences are compared. Similarity judgment algorithms come in many different forms. One angle of this article is to analyze the assumptions behind such similarity metrics by looking at different shared or non-shared properties. Among the interesting properties we note symmetry and transitivity, which are fundamental to the understanding of similarity. This angle is investigated in Section 4 and 5, looking respectively at two broad classes of mathematical models of similarity and then more closely at different similarity metrics.

As Senseval and other similar competitions need a gold standard for evaluating the competing systems, the second angle of our research looks into literature in philosophy and psychology to gain insight on the human capability in performing a similarity judgment. From the first discipline explored in Section 2, we discover that philosophers have divergent views on concept identification, ranging from scientific definitions to human perception of concepts. From the second discipline, explored in Section 3, we discover different psychological models for concept identification and implicitly concept comparison, this time ranging from continuous concepts being positioned in multi-dimensional spaces to concrete concepts being grasped as entities.

The two angles (metrics and humans) converge in the conclusion of Section 6 with general observations and future work.

2 Philosophical evidence

Children have a natural eagerness to recognize regularities in the world and to mimic the behavior of competent members of their linguistic community. It is in these words that Wittgenstein (1980) simply expresses how infants acquire the community's language. What underlies the activities surrounding a common use of language is similar to our usage of words to express something: "Consider for example the proceedings that we call *games*. I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?" (Wittgenstein, 1968: 66). Wittgenstein answers that these expressions are characterized by similarities he calls *family resemblances.*

Proceedings of the Workshop on Linguistic Distances, pages 8–15,
Sydney, July 2006.

Given that a dictionary's purpose is to define concepts, we could hope to see such family resemblances among its definitions. Contrarily to this intuition, Table 1 shows definitions and examples for a few senses of *game* in Wordnet[1], from which resemblance cannot be found in terms of common words in the definitions or examples. Nevertheless, humans are able to give different judgments of similarity between different senses of the word *game*. For example, similarity between sense 1 and sense 3 is intuitively larger than between sense 1 and sense 4.

Table 1: Some senses of *game* in Wordnet

	Definition + *Example*
1	A single play of a sport or other contest. *The game lasted two hours.*
2	A contest with rules to determine a winner. *You need four people to play this game.*
3	The game equipment needed in order to play a particular game. *The child received several games for his birthday.*
4	Your occupation or line of work *He's in the plumbing game.*
5	A secret scheme to do something (especially something underhand or illegal). […] *I saw through his little game from the start.*

Before being tempted to call up gigabytes of corpus evidence data and computational strength to help us identify the family of resemblance emerging here, let us further look at the nature of that notion from a philosophical point of view. Possible senses of individual things could be traced back to Aristotle's work and identified "without qualification" as the primary substance of a thing (Cassam, 1986). What accounts for the substance of an object, for Aristotle, was the thing itself, namely its essence. Taking a slightly different view on the notion of family of objects, Putnam (1977) instead pursues a quest for *natural kinds* and according to him, the distinguishing characteristics that "hold together" natural kinds are the "core facts […] conveying the use of words of that kind" (Putnam, 1977: 118). Putnam disagrees with any analytical approaches sustaining that the meaning of a word X is given by a conjunction of properties $P = \{P_1, P_2,\ldots P_n\}$ in such a way that P is the essence of X. The problem is that a "natural kind may have *abnormal members*" (Putnam, 1977: 103). For instance, normal lemons have a yellow peel but let's suppose in accordance with Putnam, that a new environmental condition makes lemon peel become blue. An analytical view will be unable to state which one amongst the yellow or the blue ones is now the normal member of the natural class of lemons. Putnam rather relies on a "scientific theory construction" to define what an object of natural kind is, and therefore, does not see that dictionaries "are *cluttered up* […] with pieces of empirical information" (Putnam, 1977: 118) as a defect to convey core facts about a natural class.

In contrast to Putnam, Fodor (1998) is a virulent opponent to a mind-independent similarity semantics subject to scientific discoveries. With his ostentatious *doorknob* example, Fodor shows that there is not any natural kind, hidden essence or peculiar structure that makes a doorknob *a doorknob*. "No doubt, some engineer might construct a counter-example–a mindless doorknob detector; and we might even come to rely on such a thing when groping for a doorknob in the dark" (Fodor, 1998: 147). However, the construct will have to be done on what strikes us as *doorknobhood* or satisfying the *doorknob stereotype*, i.e. "the gadget would have to be calibrated to us since there is nothing else in nature that responds selectively to doorknobs" (Fodor, 1998: 147). According to Fodor, our capacity to acquire the concept of *doorknob* involves a similarity metric, and it is the human innate capacity to determine the concepts similar to *doorknob* that allow the characterization of *doorknobhood*. Therefore, Fodor states that the meaning of concepts is mind-dependent and that individuation is not intractable since members of a language community, although experiencing diverse forms of a concept will tend to acquire similar stereotypes of such a concept.

This brief exploration into philosophical approaches for concept representation and delimitation can inform us on the establishment of a gold standard by humans for the word sense disambiguation (WSD) task. In fact, the adherence to one model rather than another has an impact on who should be performing the evaluation[2]. Senseval-2 was in line with Putnam's view of 'division of linguistic labour' by relying on lexicographers' judgments to build a gold standard (Kilgarrif, 1998). On the other hand, Senseval-3 collected data via Open-Mind Initiative[3], which was much more in line with Fodor's view that any common people can use their own *similarity*

[1] See http://wordnet.princeton.edu/

[2] The evaluation consists in performing sense tagging of word occurrences in context.

[3] See http://www.openmind.org/, a web site where anyone can perform the sense tagging "games".

metric to disambiguate polysemous terms. Interestingly, a recent empirical study (Murray and Green 2004) showed how judgments by ordinary people were consistent among themselves but different from the one of lexicographers. It is important to decide who the best judges are; a decision which can certainly be based on the foreseen application, but also, as we suggest here, on some theoretical grounds.

3 Psychological Evidence

We pursue our quest for insights in the establishment of gold standards by humans for the WSD task, now trying to answer the "how" question rather then the "who" question. Indeed, Fodor's view might influence us in deciding that non-experts can perform similarity judgments, but this does not tell us how these judgments should be performed. Different psychological models will give possible answers. In fact, similarity judgments have been largely studied by experimental psychologists and distinctive theories give some evidence about the existence of a human internal cognitive mechanism for such judgments. In this section, we present three approaches: *subjective scaling* and *objective scaling* (Voinov, 2002), and *semantic differential* (Osgood et al. 1957).

3.1 Subjective Scaling

In *subjective scaling* (Voinov, 2002), the subjective human judgment is considered as a convenient raw material to make comparison between empirical studies of similarity. Subjects are asked to point out the "similarities among *n* objects of interest – whether concepts, persons, traits, symptoms, cultures or species" (Shepard, 1974: 373). Then the similarity judgments are represented in an $n \times n$ matrix of objects by a multidimensional scaling (MDS) of the distance between each object. Equation 1 shows the evaluation of similarity, where $d(x_{ik}, x_{jk})$ stands for the distance between objects x_i and x_j on stimulus (dimension) k and w_k is the psychological salience of that stimulus *k*:

$$D(x_i, x_j) = \sum_{k=1}^{m} w_k (d(x_{ik}, x_{jk})) . \quad (1)$$

Shepard's MDS theory assumes that a monotonic transformation should be done from a nonmetric psychological salience of a stimulus to a metric space model. By definition, the resulting metric function over a set X should fullfill the following conditions:

$\forall x, y, z \in X$:

1. $d(x, y) \geq d(x, x) = 0$ (minimality),
2. $d(x, y) = d(y, x)$ (symmetry),
3. $d(x, y) \geq d(x, z) + d(z, y)$ (triangle ineq.).

Accordingly to Shepard (1974), the distance in equation (1) can be computed with different metrics. Some of these metrics are given in Lebart and Rajman (2000). The *Euclidean metric* is the best known:

$$d_E(x_i, x_j) = \left(\sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2 \right)^{1/2} . \quad (2)$$

The *city block metric* is another one:

$$d_C(x_i, x_j) = \sum_{k=1}^{m} w_k \left| x_{ik} - x_{jk} \right| . \quad (3)$$

Another yet is the *Minkowski metric*:

$$d_N(x_i, x_j) = \sum_{k=1}^{m} w_k \left((x_{ik} - x_{jk})^n \right)^{\frac{1}{n}} . \quad (4)$$

There is a main concern with the MDS model. Tversky (1977) criticized the adequacy of the metric distance functions as he showed that the three conditions of minimality, symmetry and triangle inequality are sometimes empirically violated. For instance, Tversky and Gati showed empirically that assessment of the similarity between pairs of countries was asymetric when they asked for "the degree to which Red China is similar to North Korea" (1978: 87) and in the reverse order, i.e. similarity between North Korea and Red China.

3.2 Objective Scaling

The second approach is called *objective scaling* by Voinov "though this term is not widely accepted" (Voinov, 2002). According to him, the objectivity of the method comes from the fact that similarity measures are calculated from the ratio of objective features that describe objects under analysis. So, subjects are asked to make qualitative judgments on common or distinctive features of objects and the comparison is then made by any distance axioms. Tversky's (1977) *contrast model* (CM) is the best known formalization of this approach. In his model, the measure of similarity is computed by:

$$S(A,B) = \alpha f(A \cap B) - \beta f(A - B) - \chi f(B - A) \quad (5)$$

where $f(A \cap B)$ represents a function of the common features of both entities *A* and *B*, $f(A-B)$ is the function of the features belonging to *A* but not *B*, $f(B-A)$ is the function of the features belonging to *B* but not *A* and α, β, χ are their respective weighting parameters. Equation (5) is the *matching* axiom of the CM. A second fundamental property of that model is given by the axiom of *monotonicity*:

$$S(A,B) \geq S(A,C) \quad (6)$$

If $A \cap C \subset A \cap B$, $A-B \subset A-C$, and $B-A \subset C-A$, then (6) is satisfied. With these two axioms (5-6), Tversky (1977) defined the basis of what he called the *matching function* using the theoretical notion of feature sets rather then the geometric concept of similarity distance. Interesting empirical studies followed this research on CM and aimed at finding the correlation between human judgments of similarity and difference. Although some results show a correlation between these judgments, there is limitation to their complementarity: "the relative weights of the common and distinctive features vary with the nature of the task and support the focusing hypothesis that people attend more to the common features in judgments of similarity than in judgments of the difference" (Tverski and Gati, 1978: 84). Later on, Medin et al. (1990) also reported cases when judgments of similarity and difference are not inverses: first, when entities differ in their number of features, and second when similarity/difference judgments involve distinction of both attributes and relations. "Although sameness judgments are typically described as more global or non-analytic than difference judgments, an alternative possibility is that they focus on relations rather than attributes" (Medin et al., 1990: 68).

3.3 Semantic Differential

One standard psycholinguistic method to measure the similarity of meaning combines the use of *subjective scaling* transposed in a semantic space. One well-known method is *Semantic Differential* (SD) developed by Osgood et al. (1957).

The SD methodology measures the meanings that individual subjects grant to words and concepts according to a series of factor analyses. These factor analyses are bipolar adjectives put at each end of a *Likert scale* (Likert, 1932) devised to rate the individual reaction to the contrasted stimulus. For instance, the SD of a concept can be rated with two stimuli of goodness and temperature:

$$Good \quad \underline{\quad}_{3} : \underline{\quad}_{2} : \overset{\times}{\underline{\quad}}_{1} : \underline{\quad}_{0} : \underline{\quad}_{1} : \underline{\quad}_{2} : \underline{\quad}_{3} \quad Bad$$

$$Cold \quad \underline{\quad}_{3} : \underline{\quad}_{2} : \underline{\quad}_{1} : \underline{\quad}_{0} : \overset{\times}{\underline{\quad}}_{1} : \underline{\quad}_{2} : \underline{\quad}_{3} \quad Hot$$

If the subject feels that the observed concept is neutral with regards to the polar terms, his check-mark should be at the position 0. In our example, the mark on the *good-bad* scale being at the 1 on the left side of the neutral point 0, the judgment means *slighthy good*. Positions 2 and 3 on that same side would be respectively *quite good* and *extremely good*. A similar analysis applies for the *cold-hot* scale shown.

The theoretical background of that methodology, which tries to standardize across subjects the meaning of the same linguistic stimulus, relies on psychological research on synestesia. Simply explained, synestesia is similar to a double reaction to a stimulus. For example, when presented with images of concepts, subjects do not only have a spontaneous reaction to the images, but they are also able to characterize the associated concept in terms of almost any bipolar adjective pairs (hot-cold, pleasant-unpleasant, simple-complex, vague-precise, dull-sharp, static-dynamic, sweet-bitter, emotional-rational, etc.). According to Osgood et al. "the imagery found in synesthesia is intimately tied up with language metaphor, and both represent *semantic* relations" (1957: 23).

In SD, bipolar adjectives used in succession can mediate a generalization to the meaning of a sign, as uncertainty on each scale is reduced with the successive process of elicitation. By postulating representation in a semantic space, each orthogonal axis of selection produces a semantic differentiation when the subjects rate the semantic alternatives on a bipolar scale. Although that space could be multidimensional, empirical studies (Osgood et al., 1957) on factor analysis showed stability and relative importance of three particular dimensions labeled as Evaluation, Potency, and Activity (EPA). We refer the reader to Osgood et al. (1957) for further explanation on these EPA dimensions.

3.4 WSD and human judgments

Table 2 emphasizes commonalities and differences between the three psychological models explored.

Table 2 – Psychological Models

	Continuous	Prede-fined traits	Similarity/ Difference
MDS	Yes	Yes	No
CM	No	Yes	Yes
SD	No	No	Possible

In Table 2, we show that both MDS (Shepard, 1974) and CM (Tversky, 1977) rely on a set of predefined traits. This is a major problem, as it leads to the necessity of defining in advance such a set of traits on which to judge similarity between objects. On the other hand, SD (Osgood et al. 1957), although using a few bipolar scales for positioning concepts, argues that these scales are not concept-dependent, but rather they can be used for grasping the meaning of all concepts. A second major difference highlighted in Table 2 is that MDS is the only approach looking at continuous perceptual dimensions of stimulus, contrarily to CM in which the scaling procedes with discrete conceptual traits, and even more in opposition to SD which considers entities as primitives. Finally, Table 2 shows the interesting observation brought forth by Tversky and later empirical studies of Medin et al. (1980) of the non-equivalence between the notion of similarity and difference.

Coming back to the question of "how" human evaluation could be performed to provide a gold standard for the WSD task, considering the pros and cons of the different models lead us to suggest a particular strategy of sense attribution. Combining the similarity/difference of Tversky with the successive elucidation of Osgood et al., two bipolar Likert scales could be used to delimit a similarity concept: a resembling axis and a contrasting axis. In this approach, the similarity concept still stays general, avoiding the problems of finding specific traits for each instance on which to have a judgment.

Already in the empirical studies of Murray and Green (2004), a Likert scale is used, but on an "applying" axis. Subjects are asked for each definition of a word to decide whether it "applies perfectly" or rather "barely applies" to a context containing the word. The choice of such an axis has limitations in its applicability for mapping senses on examples. More general resembling and contrasting axis would allow for similarity judgments on any statements whether they are two sense definitions, two examples or a sense definition with an example.

4 Mathematical Models of Similarity

Logic and mathematics are extremely prolific in similarity measurement models. According to Dubois et al (1997), they are used for cognitive tasks like classification, case-based reasoning and interpolation. In the present study, we restrict our investigation to the classification task as representative on the unsupervised WSD task. The other approaches are inferential strategies, using already solved problems to extrapolate or interpolate solutions to new problems. Those would be appropriate for WSD in a supervised context (provided training data), but due to space constraints, we postpone discussion of those models to a later study. Our present analysis divides classification models into two criteria: the *cardinality of sets* and the *proximity-based* similarity measures.

4.1 Cardinality of sets

In line with De Baets et al. (2001), similarity measures can be investigated under a rational *cardinality*-based criterion of *sets*. In an extensive study of 28 similarity measures for ordinary sets, this research showed that measures can be classified on the basis of only a few properties. They proposed at first to build the class of cardinality-based similarity measures from one generic formula:

$$S(X,Y) = \frac{w\alpha_{X,Y} + x\beta_{X,Y} + y\chi_{X,Y} + z\delta_{X,Y}}{w'\alpha_{X,Y} + x'\beta_{X,Y} + y'\chi_{X,Y} + z'\delta_{X,Y}}, \quad (8)$$

where $\alpha_{X,Y} = \min\{\#(X-Y), \#(Y-X)\}$, $\beta_{X,Y} = \max\{\#(X-Y), \#(Y-X)\}$, $\chi_{X,Y} = \#(X \bigcap Y)$ and $\delta_{X,Y} = \#(X \bigcup Y)^c$, and all w, x, y, z, w', x', y', $z' \in \{0,1\}$. It follows that $\#(X \bigcap Y)$ is the number of couples (1,1) and $X - Y$ denotes the sets difference $(X-Y) = (X \bigcap Y^c)$.

The classification of these 28 similarity measures (which can all be linked to the general formula) becomes possible by borrowing from the framework of fuzzy sets the concepts of *T* for t-norm (*fuzzy intersection*) operators and *T-equivalence* for the property of *T*-indistinguishability (De Baets et al., 2001). So, a typical measure *M* of *T-equivalence* under the universe U must satisfy the following conditions for any (*x*, *y*, *z*) $\in U$: (i) $M(x,x) = 1$ (reflexivity); (ii) $M(x,y) = M(y,x)$ (Symmetry);

(iii) $T(M(x,y),M(y,z)) \leq M(x,z)$ (*T*-transitivity).

All 28 measures show reflexivity and symmetry but they vary on the type of transitivity they achieve. In fact, studying boundary and monotonicity behavior of the different measures, De Baets et al. (2001) group them under four types corresponding to four different formulas of fuzzy intersections (t-norms): the standard intersection $Z(a,b) = \min(a,b)$, the Lukasiewicz t-norm $L(a,b) = \max(0, a+b-1)$, the algebraic product $P(a,b) = ab$ and the drastic intersection $D(a,b) = (a$ when $b = 1$, b when $a = 1$ and 0 otherwise). We refer the reader to De Baets et al. (2001) to get the full scope of their results. Accordingly, Jaccard's coefficient *J* (equation 9) and Russel-Rao's coefficient *R* (equation 10) are both, for example, *L*-transivive (Lukasiewicz' type):

$$S_J(X,Y) = \frac{\#(X \bigcap Y)}{\#(X \bigcup Y)} \quad (9)$$

$$S_R(X,Y) = \frac{\#(X \bigcap Y)}{n} \quad . \quad (10)$$

On the other hand, the overlapping coefficient *O* (equation 11) is not even *D*-transitive, knowing that *D* is the lower transitive condition $(D \leq L \leq P \leq Z)$ in the framework:

$$S_O(X,Y) = \frac{\#(X \bigcap Y)}{\min(\#X, \#Y)} \quad . \quad (11)$$

4.2 Proximity-based

Following our second criterion of classification, mathematics also uses diverse *proximity-based* similarity measures. We subdivide these mathematical measures into three groups: the distance model, the probabilistic model, and the angular coefficients. The first one, the distance model, overlaps in part with the subjective scaling of similarity as presented in the psychological approaches (section 3.1). The mathematical model is the same with a metric of distance $d(x,y)$ computed between the objects in a space. Algorithms like formulae (2), (3) and (4) of section 3.1 are amongst the *proximity-based* similarity measures.

Second, the probabilistic model is based on the statistical analysis of objects and their attributes in a data space. Lebart & Rajman (2000) gave many examples of that kind of proximity measures, such as the Kullback-Leiber distance D_K between two documents *A* and *B*, given the probability distribution $P = \{p_1, p_2, ..., p_n\}$:

$$D_K(A,B) = \sum_{p_{ak} \times p_{bk} \neq 0} (p_{ak} - p_{bk})(\log p_{ak} - \log p_{bk}) \quad (12)$$

The third mathematical model is also a metric space model but it uses angular measures between vectors of features to determine the similarity between objects. A well-known measure from that group is the cosine-correlation:

$$S_C(x,y) = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\left[\sum_{k=1}^{n} x_k^2\right]\left[\sum_{k=1}^{n} y_k^2\right]}} \quad . \quad (13)$$

Although conditions applying on proximity-based measures are shortly described in Cross and Sudkamp (2002) and Miyamoto (1990) for fuzzy sets, we are not aware of an extensive research such as the one by De Baets et al. (2001), presented in section 4.1, for classifying cardinality of sets types. We make such an attempt in the following section.

5 Analysis of similarity metrics

In this section, we perform a classification and analysis exercise for similarity measure[4], possibly used for WSD, but more generally used in any task where similarity between words is required. Table 3 shows the measures classified in the four categories of the mathematical model presented in section 4: measures of cardinality (Card), of distance (Dist), of probability (Prob) and of angle (Ang).

We sustain that these groupings can be further justified based on two criteria: the psychological model of meaning (Table 2) and the typical properties of the classes (Table 4). The first criterion refers to the representation of concepts distinguishing between the dense-state and the discrete-state[5] of concept (meaning) attributes. That psychological distinction is helpful to categorize some metrics, like Gotoh, which seems hybrid (Card and Dist). In such a metric, the penalty for the gap between two concepts applies on the defect of the dense-state, such as for a blurred im-

[4] We use the list of the following web page: http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#sellers

[5] This differentiation is based on Tenenbaum's (1996) idea that MDS better suits continuous perceptual domains and set-theoretic accommodate discrete features like in the CM.

age rather then the absence of the discrete-state, i.e. of a feature; it is therefore classified in the Dist category.

Table 3: Classification of Similarity Metrics

Metric	Card	Dist	Prob	Ang
Hamming distance		X		
Levenshtein distance		X		
Needleman-Wunch		X		
Smith-Waterman		X		
Gotoh distance		X		
Block distance		X		
Monge Elkan dist.		X		
Jaro distance			X	
Jaro Winkler			X	
SoundEx distance			X	
Matching coefficient	X			
Dice's coefficient	X			
Jaccard similarity	X			
Overlap coefficient	X			
Euclidean distance		X		
Cosine similarity				X
Variational distance			X	
Hellinger distance			X	
Information radius			X	
Harmonic mean			X	
Skew divergence			X	
Confusion probability			X	
Tau			X	
Fellegi & Sunters			X	
TFIDF				X
FastA			X	
BlastP			X	
Maximal matches			X	
q-gram			X	
Ukkonen algorithms			X	

The second criterion is a study on shared properties for each category of the mathematical model. Table 4 summarizes the properties using the following schema: (m) minimality, (r) reflexivity, (s) symmetry, (ti) triangle inequality, (tr) transitivity.

Table 4 – Typical Properties of Metrics

	(m)	(r)	(s)	(ti)	(tr)
Card		Yes	Yes		Yes
Dist	Yes		Yes	Yes	Possible
Prob		No	Possible		Yes
Ang	Yes		Yes		Yes

From Table 4, we see for instance that reflexivity is a basic property for cardinality measures because we wish to regularly count discrete objects in a set. On the opposite side, the minimality property is a characteristic of a distance measure, since it is noticeable by the displacement or the change, for example, in distinctive images. According to Fodor (1998), we say that statistical or probabilistic approaches exhibit several necessary and sufficient conditions for the inclusion of elements in the extension of a concept, but the dominant element, such as the pattern of comparison (in Maximal matches for instance) is anti-reflexive and asymmetric with the resulting elements. However, there is symmetry in the resultant, but there is still anti-reflexivity.

We also single out the angular metrics from distance measures even though they use a similar analysis of the qualitative variation of entities. According to Ekman & Sjöberg (1965), a method using similarity converted into cosine representation has the advantage to reveal two components of percepts, i.e. the two-dimensional vector is a modeling in magnitude and direction. Thus, angular metrics can be a means used to contrast two semantic features of entities.

5.1 A closer look at properties

Finding out that different sets of properties can serve as dividing lines between groups of metrics is interesting in itself, but does not answer the question as to which set is more appropriate than others. We do not wish to answer this question here as we believe it is application-dependent, but we do wish to emphasize that a questioning should take place before choosing a particular measure. In fact, for each property, there is an appropriate question that can be asked, as is summarized in Table 5.

Table 5 – Questioning for Measure Selection

Property	Question
Minimality	Is the minimal distance between objects the distance of an object with itself?
Symmetry	Is it true that the distance between x and y is always the same as the distance between y and x?
Triangle Inequality	Is it appropriate that a direct distance between x and z is always smaller than a composed distance from x to y and y to z?
Reflexivity	Is it true that the relation that it holds between an object and itself is always the same?
Transitivity	Is it necessarily the case that when x is similar to y and y is similar to z, that x be similar to z?

For the task of WSD investigated in this paper, we hope to open the debate as to which properties are to be taken into consideration.

6 Conclusion and future work

This paper presented some ideas from two angles of study (human and metrics) into the intricate problem of similarity judgments. A larger study

is under way on both angles. First, we suggested, based on some psychological and philosophical model analysis, a two-axis Osgood-like benchmarking approach for "ordinary human" word-sense judgments. We intend to perform an empirical experiment to validate this idea by looking at inter-judge agreement.

On the algorithm side, although the approaches based on the cardinality of sets are not central to WSD, we presented them first as we find it inspiring to see an effort of classification on those measures. We then attempted a somewhat more broad classification by emphasizing properties of different groups of similarity measures: cardinality of sets, distance, probabilistic measures and angular metrics. Although each group has a particular subset of properties, we noted that all of them share a property of transitivity. This is interestingly different from the psychological contrast model of Tversky where differences and similarities are measured differently on different criteria. We think investigations into similarity measures which reproduce such a non-transitive differentiation approach should be performed. We are on that path in our larger study. We also suggest that any proposal of a measure for a task should be preceded by a study of which properties seem adequate for such a task. We conclude by opening up the debate for the WSD task.

References

Bernard De Baets, Hans De Meyer and Helga Naessens. 2001. A class of rational cardinality-based similarity measures. *Journal of Computational and Applied Mathematics*, 132:51-69.

Quassim Cassam. 1986. Science and Essence. *Philosophy*, 61:95-107.

Valerie V. Cross and Thomas A. Sudkamp. 2002. *Similarity and Compatibility in Fuzzy Set Theory*. Heidelberg, Germany: Physica-Verlag..

Didier Dubois, Henri Prade, Francesc Esteva, Pere Garcia and Lluis Godo. 1997. A Logical Approach to Interpolation Based on Similarity Relations. *International Journal of Approximate Reasoning*, 17:1-36.

Cösta Ekman and Lennart Sjöberg. 1965. Scaling. *Annual Review of Psychology*, 16, 451-474.

Jerry A. Fodor. 1998. *Concepts. Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Adam Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12:453-472.

Ludovic Lebart and Martin Rajman. 2000. Computing Similarity in R. Dale, H. Moisl & H. Somers eds. *Handbook of Natural Language Processing*. New York: Marcel Dekker, Inc., 477-505.

Rensis Likert. 1932. A technique for the measurement of attitudes. Archives of Psychology 140, 5-53.

Douglas L. Medin, Robert L. Goldstone and Dedre Gentner. 1990. Similarity Involving Attributes and Relations: Judgments of Similarity and Difference are not Inverses. *Psychological Science*, 1(1):64-69

Rada Mihalcea and Phil Edmonds. 2004. *Proceedings of SENSEVAL-3, Association for Computational Linguistics Workshop*, Barcelona, Spain.

Sadaaki Miyamoto. 1990. *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht: Kluwer Academic Publisher.

G. Craig Murray and Rebecca Green. 2004. Lexical knowledge and human disagreement on a WSD task, *Computer Speech and Language* 18, 209-222.

Charles E. Osgood, George J. Suci and Percy H. Tannenbaum. 1957. *The measurement of meaning*. Urbana: University of Illinois Press

Judita Preiss and David Yarowsky (eds). 2001. *Proceedings of SENSEVAL-2, Association for Computational Linguistics Workshop*, Toulouse, France.

Hilary Putnam. 1977. Is Semantics Possible? in Stephen P. Schwartz ed. *Naming, Necessity, and Natural Kinds*. Ithaca and London: Cornell University Press, 102-118.

Roger N. Shepard. 1974. Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39(4):373-421.

Joshua B. Tenenbaum. 1996. Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (Eds), *Advances in neural information processing systems*, (Vol. 8, pp. 3-9), Cambridge, MA: MIT Press.

Amos Tversky. 1977. Features of Similarity. *Psychological Review*, 84, 79-98.

Amos Tversky and Itamar Gati. 1978. Studies of Similarity in E. Rosch & B. B. Lloyd eds. *Cognition and Categorization*. New York: John Wiley & Sons, Inc., 79-98.

Alexander V. Voinov. 2002. The Role of Similarity Judgment in Intuitive Problem Solving and its Modeling in a Sheaf-Theoretic Framework. *Proceedings of the 1st Int. Conf. on FSKD'02*, 1:753-757.

Ludwig Wittgenstein. 1968. *Philosophical Investigations*. Oxford: Basil Blackwell.

Ludwig Wittgenstein. 1980. *Remarks on the Philosophy of Psychology*. Chicago: University of Chicago Press; Oxford: Basil Blackwell.