Technical Note

A Discussion on the Property of the Semantic Space in the SD-Form Semantics Model

Masahiro Wakiyama,[†1] Shouta Yoshihara,[†2] Hideki Noda,[†3]
Koichi Nozaki[†4] and Eiji Kawaguchi[†3]

The SD-Form Semantics Model, proposed by the authors, is a framework to deal with semantic data for natural language processing. The most important idea in this model is an introduction of meaning description language (SD-Form), as well as a proposal of a semantic-difference score between two language expressions. In this article we discuss the triangular property of a semantic space defined by the SD-Form. We also demonstrate that an inductive inference from many facts is not necessarily the same as the one from two facts in natural language expression.

1. Introduction

Many researchers have been working on the semantics of natural language hoping for an implementation of machine intelligence. Researchers in many different disciplines often share the same topics and address a problem by using common phrases. But they seek for different methodologies from a different point of view. So, the semantics is an interdisciplinary topic, yet it has different approaches in each discipline [2].

The authors previously proposed a semantics model by introducing a formal language (SD-Form) [3] as an interlingua. In the present article we discuss the triangular property of a semantic space with reference to our model.

In Section 2 we discuss the property of a semantic space in human brain and relate it to our model. In Section 3 we will demonstrate that there is some discrepancy between a two-facts based induction and a multi-facts based induction. Finally, in Section 4 we summarize the discussion.

2. Triangular Property of a Semantic Space in Natural Language

In this section we discuss a general problem with respect to a semantic space in natural language. Our concern is whether a semantic space satisfies the conditions of metric space or not.

Mathematically, a function $f(a, b)$ is called a metric if the following properties are satisfied:

1) $f(a, a) = 0$ (zero property),
2) $f(a, b) = f(b, a)$ (symmetric property),
3) $f(a, b) \geq 0$ (positive property),
4) $f(a, b) + f(b, c) \geq f(a, c)$ (triangular property).

Rada, et al. [1] discussed Properties 2) and 4) in depth regarding a conceptual space in natural language.

In our model [3], ***DIFF*** satisfies conditions 1), 2), and 3). While 4) is the condition we did not incorporate in the model.

A semantic space in most traditional models was introduced as a set of simple words which are mostly nouns. In such models the triangular property was appreciated from an information retrieving point of view because this property guarantees a retrieved result from a step-by-step method is close to the result from a direct access method. However, we think such a property is not true in a more generalized space. In the following examples we see some English expressions which do not comply with Property 4).

⟨Example 2-1⟩

⟨Group-1⟩
1-a: It is hot now.
1-b: It is sunny now.
1-c: It is cold now.

⟨Group-2⟩
2-a: He plays golf very well.
2-b: He is a golf instructor.
2-c: He is a sports instructor.

†1 Department of Control & Information Systems Engineering, Kitakyushu National College of Technology
†2 Department of British-American Culture Studies, Junshin Junior College
†3 Department of Electrical, Electronic & Computer Engineering, Kyushu Institute of Technology
†4 Information Science Center, Nagasaki University

Table 1 A semantic distance rating test.

Group-1	a-b	b-c	c-a	
A	0	2	9	*
B	3	4	9	*
C	2	6	9	*
D	1	5	4	
E	1	4	3	
Group-2	**a-b**	**b-c**	**c-a**	
A	1	1	3	*
B	9	6	9	
C	4	2	7	*
D	1	5	6	
E	1	4	4	*
Group-3	**a-b**	**b-c**	**c-a**	
A	9	0	9	
B	7	2	7	
C	8	8	1	
D	2	9	7	
E	3	6	5	

⟨Group-3⟩

3-a: What are you doing?

3-b: How are you doing?

3-c: How are you?

With these sentences as examples we asked 5 students ($\mathbf{A}, \mathbf{B}, \ldots, \mathbf{E}$) to rate a semantic distance score for three combinations of the sentences in each group. We suggested the semantic distance score should range from 0 to 9, with 0 being identical with each other, while 9 being entirely different from the other. The subjects **A**, **B** and **C** were native English speakers, while **D** and **E** were non-natives. The results are shown in **Table 1**.

The scores in this list can be interpreted as the values of $f(a,b)$, $f(b,c)$, and $f(c,a)$ for each group. The scores with $*$ mark do not satisfy the triangular property.

As we see here, for Group-1, subjects **A** and **B** gave scores that apparently violate the triangular property. In particular, they rated "It is sunny now" and "It is cold now" as quite similar in meaning.

This is because their home town is in the far north of the US where they had many cold days with sunshine in winter.

This example shows that semantic distance in natural language depends on the background knowledge (or personal experiences), which may lead to violation of the triangular property of a metric space.

The SD-Form expressions of group-1 can be as follows:

1-a: $[s(\text{WEATH-COND/NOW}), v(\text{HOT})]$,

1-b: $[s(\text{WEATH-COND/NOW}), v(\text{SUNSHINE})]$,

1-c: $[s(\text{WEATH-COND/NOW}), v(\text{COLD})]$.

The background knowledge of the subjects **A** and **B** might be something like the following:

K1: (HOT) *incl* (SUNSHINE),

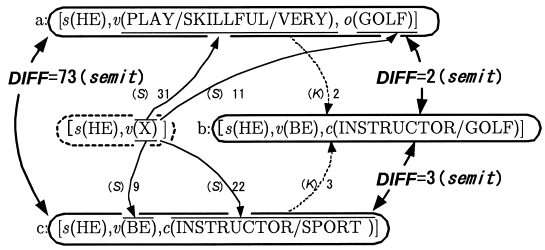

Fig. 1 Semantic difference scores for group-2.

K2: (COLD) *incl* (SUNSHINE).

For this knowledge combination, our experimental system (**SDENV-2**)[3] evaluates the following semantic distance scores:

$$\boldsymbol{DIFF}(a,b)=3,\ \boldsymbol{DIFF}(b,c)=3,$$
$$\boldsymbol{DIFF}(c,a)=18$$

which also violates the metric space condition.

For the sentences in group-2, $\boldsymbol{DIFF}$ computation in **SDENV-2** is as follows:

⟨**Example 2-2**⟩

SD-Forms for each English expression:

2-a: $[s(\text{HE}), v(\text{PLAY/SKILLFUL/VERY}), o(\text{GOLF})]$,

2-b: $[s(\text{HE}), v(\text{BE}), c(\text{INSTRUCTOR/GOLF})]$,

2-c: $[s(\text{HE}), v(\text{BE}), c(\text{INSTRUCTOR/SPORT})]$.

Knowledge:

K1: $(assu([s(\text{X}), v(\text{BE}), c(\text{INSTRUCTOR/Y})]))$
$caus([s(\text{X}), v(\text{PLAY/SKILLFUL/VERY}), o(\text{Y})])$
(If X is an instructor of Y, then X plays Y very well.)

K2: $(\text{SPORT})incl(\text{GOLF})$
(Sports include golf.)

Instantiated Rule:

K3: $(assu([s(\text{HE}), v(\text{BE}), c(\text{INSTRUCTOR/GOLF})]))$
$caus([s(\text{HE}), v(\text{PLAY/SKILLFUL/VERY}), o(\text{GOLF})])$

The semantic difference scores are as follows (c.f, **Fig. 1**):

$$\boldsymbol{DIFF}(a,b) = 2,$$
$$\boldsymbol{DIFF}(b,c) = 3,$$
$$\boldsymbol{DIFF}(a,c) = 73.$$

These scores also violate the metric space condition.

Generally speaking, each pair of language expressions has a connotation that implicitly activates a topic world. In some world two expressions are very similar, but in other worlds

they are very different. This is because the semantic difference between two expressions depends on the topic world. Therefore, we think that the triangular property should not be introduced in our model. Some people may object to this view, on the grounds that the topic world is common to the speaker and the listener when they are talking. Certainly, it is true that we talk to each other in the belief that we are discussing the same topic. However, this is not always true. Moreover, even if the topic world is common to every expression, we think that the triangular property does not necessarily hold true. Therefore, we omitted the triangular property in the SD-Form Semantics Model.

We do not claim that our model explains the human semantic distance scores. Rather, we claim that the SD-Form semantics model can adapt to a knowledge-depending distance evaluation.

3. Induction from Many Given Statements

In ordinary circumstances, we feel that an induction from two statements is rather easy, while an induction from many statements is quite difficult. We will interpret this situation by using our model.

A conclusion from many statements is equivalent to an induction from many facts. In the SD-Form semantics model it is equivalent to detecting the nearest common ancestor (***NCOA***) of many concepts.

However, we cannot easily find out such ***NCOA*** by using the ***NCOA***-detecting algorithm for a concept pair (i.e., $\boldsymbol{G}(M, D_1, D_0, D_2, n_0)$ in Ref. 3)). Let us consider the following example:

⟨Example 3-1⟩

Knowledge:

(HUMAN)*incl*([AMERICAN,WENDY]),
(STUDENT)*incl*([BILL,WENDY]),
(FAMILY/CLARK)*incl*([ALICE,BILL]),
(GIRL)*incl*([ALICE,WENDY]),
(AMERICAN)*incl*(BILL),
(AMERICAN/PRETTY)*incl*(ALICE).

A concept triple:

(WENDY, ALICE, BILL).

In this circumstance we see in **Fig. 2** that HUMAN is the ***NCOA*** of this concept triple, because

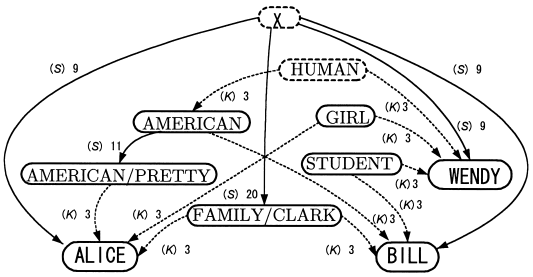

Fig. 2 *NCOA for a concept triple.*

$$
\begin{aligned}
&\boldsymbol{ELAB}(\text{HUMAN,ALICE})\\
&\quad = \boldsymbol{ELAB}_{\boldsymbol{know}}(\text{HUMAN,AMERICAN})\\
&\quad +\boldsymbol{ELAB}_{\boldsymbol{synt}}(\text{AMERICAN,}\\
&\qquad\qquad \text{AMERICAN/PRETTY})\\
&\quad +\boldsymbol{ELAB}_{\boldsymbol{know}}(\text{AMERICAN/PRETTY,}\\
&\qquad\qquad \text{ALICE})\\
&\quad = 17,\\
&\boldsymbol{ELAB}(\text{HUMAN,BILL})\\
&\quad = \boldsymbol{ELAB}_{\boldsymbol{know}}(\text{HUMAN,AMERICAN})\\
&\quad +\boldsymbol{ELAB}_{\boldsymbol{know}}(\text{AMERICAN,BILL})\\
&\quad = 6,\\
&\boldsymbol{ELAB}(\text{HUMAN,WENDY})\\
&\quad = \boldsymbol{ELAB}_{\boldsymbol{know}}(\text{HUMAN,WENDY})\\
&\quad = 3.
\end{aligned}
$$

Therefore, the total score of the elaboration from HUMAN to each of the given concepts is 26. However, HUMAN cannot be detected by any combination of $\boldsymbol{G}(M, D_1, D_0, D_2, n_0)$ algorithms. We can detect X as a common ancestor of three concepts (ALICE, BILL, WENDY) by using

$$\boldsymbol{G}(1, \text{ALICE, FAMILY/CLARK, BILL}, 6),$$
$$\boldsymbol{G}(1, \text{FAMILY/CLARK, X, WENDY}, 29)$$

steps. This X gives us

$$
\begin{aligned}
&\boldsymbol{ELAB}_{\boldsymbol{synt}}(\text{X, ALICE})\\
&+\boldsymbol{ELAB}_{\boldsymbol{synt}}(\text{X, BILL})\\
&+\boldsymbol{ELAB}_{\boldsymbol{synt}}(\text{X,WENDY})\\
&= 27.
\end{aligned}
$$

The result is a little larger than 26, which indicates that X is not the nearest common ancestor for this triple.

This example tells us that finding the ***NCOA*** for a concept triple is different from detecting the ***NCOA*** by combining pair-wise ***NCOA*** calculations. This may explain the difficulty of induction from many given facts in the human brain. It is not very difficult for our model to expand $\boldsymbol{G}(M, D_1, D_0, D_2, n_0)$ from a "pair-wise algorithm" into a "triple-wise algorithm"

by analyzing the triple combinations of (S)'s and (K)'s. Actually, our latest version of the experimental system (**SDENV-3**) is equipped with such an algorithm.

Let us now take a more difficult example of induction from a triple statement.

⟨Example 3-2⟩

A triple statements in English:

(S1) Tom likes watching boxing matches on TV.
(S2) Heather plays tennis every week.
(S3) Rick hopes to learn canoeing.

Background knowledge in English:

(F1) Tom once studied abroad in England.
(F2) Heather once lived in Paris.
(F3) Rick joined in a package tour to Rome last year.
(F4) Tom, Heather and Rick are Americans.
(F5) Paris is a part of France.
(F6) Rome is a part of Italy.
(F7) France is a part of Europe.
(F8) England is a part of Europe.
(F9) Italy is a part of Europe.
(F10) Sports include boxing, tennis and canoeing.
(F11) "Every week" is a kind of "to be regularly."
(F12) A package tour is a kind of tour.
(R1) If X likes to watch some sport Y on TV, then X likes Y.
(R2) If X does a sport Y regularly, then X likes Y.
(R3) If X wants to learn Y, then X likes Y.
(R4) If X once studied abroad in Y, then X once lived there.
(R5) If X once lived in Y, then X has once been in Y.
(R6) If X traveled to Y, then X has been in Y.
(R7) If X once visited and X is a Y, then X is a Y who once visited Z.

The conclusion **SDENV-3** can detect:

"Some Americans who once visited Europe like some sport."

(This is a translation of the detected SD-Form.)

We may also be able to implement an ***NCOA*** algorithm for a concept quartet. We are very sure that such a system is much more powerful than the human brain in inducting a conclusion from four facts. However, we do not yet have any efficient algorithms.

4. Conclusions

In this article, the authors have discussed the triangular property of a semantic space in natural language, and given several examples that violate such a property. As for the induction from given facts, they exemplified a pair-wise induction cannot detect a right conclusion from a given fact triple. The authors claim that the SD-Form model can formalize these aspects of natural-language semantics in a general way.

References

1) Rada, R., Mili, H., Bicknell, E. and Blettner, M.: Development and application of a metric on semantic nets. *IEEE Trans. System, Man, and Cybernetics*, Vol.19, pp.17–30 (1989).
2) Shapiro, S.C.: *Encyclopedia of Computer Science*, Second Edition, John Wiley & Sons, New York (1992).
3) Wakiyama, M., Noda, H., Nozaki, K. and Kawaguchi, E.: Computation Algorithm of Semantic Difference Measure in the SD-Form Semantics Model, *Trans. IPS Japan*, Vol.40, No.3, pp.1065–1079 (1999).

(Received July 8, 1999)
(Accepted November 4, 1999)