IEICE TRANS. INF. & SYST., VOL.E82–D, NO.2 FEBRUARY 1999

LETTER

Viewpoint-Based Similarity Discernment on SNAP

Takashi YUKAWA[†], *Member*, **Sanda M. HARABAGIU**[††], *and* **Dan I. MOLDOVAN**[††], *Nonmembers*

SUMMARY This paper presents an algorithm for viewpoint-based similarity discernment of linguistic concepts on Semantic Network Array Processor (SNAP). The viewpoint-based similarity discernment plays a key role in retrieving similar propositions. This is useful for advanced knowledge processing areas such as analogical reasoning and case-based reasoning. The algorithm assumes that a knowledge base is constructed for SNAP, based on information acquired from the WordNet linguistic database. The algorithm identifies paths on the knowledge base between each given concept and a given viewpoint concept, then computes a similarity degree between the two concepts based on the number of nodes shared by the paths. A small scale knowledge base was constructed and an experiment was conducted on a SNAP simulator that demonstrated the feasibility of this algorithm. Because of SNAP's scalability, the algorithm is expected to work similarly on a large scale knowledge base.

key words: *semantic network, marker propagation, analogy, inference system, parallel processing, natural language processing*

1. Introduction

Retrieving semantically related propositions or clauses is important for advanced knowledge processing areas, such as analogical reasoning and case-based reasoning. A framework which aims to achieve robust and versatile similitude retrieval that takes advantage of linguistic aspects of propositions has been proposed [1]. In that approach, which is also similar to the approach we use in this paper, the similar propositions or clauses for a given proposition vary depending on the viewpoint, which may be situation, context and intention. For instance, the clause "driving a car" is semantically closer to the clause "riding a horse" than clause "skiing" under intention of "travel event." On the other hand, giving "sport event" as intention, the clause "skiing" becomes more similar to the clause "riding a horse." Formally, this problem is defined as: given a key concept C_k, two or more candidate concepts $C_a, C_b, \ldots$, and a viewpoint concept C_v, determine the concept that is more similar to the key concept C_k under the viewpoint C_v. We represent the problem as $SD\{C_v, C_k, (C_a, C_b, \ldots)\}$. Figure 1 illustrates the problem. Concept C_a is more similar to C_k from viewpoint C_{v1} than any other concepts, and C_b is more similar from viewpoint C_{v2} than any other concepts.

For a large knowledge base, this method requires a huge volume of processing. Semantic Network Array Processor (SNAP) [3] seems to be suitable for solving the problem because it exploits the natural parallelism. It uses a marker propagation and a scalable architecture. In this paper, we propose an algorithm for computing the viewpoint-based similarity discernment for linguistic concepts on SNAP. In addition, we describe a small scale knowledge base that was constructed and an experiment on a SNAP simulator that demonstrates the feasibility of the algorithm.

2. SNAP Architecture and WordNet

SNAP is a highly parallel architecture optimized for semantic network processing with marker-propagation mechanism [3]. The knowledge is represented in a form of a semantic network and the knowledge base is distributed among the elements of the SNAP array. Reasoning on SNAP relies on marker propagation over distributed semantic networks. Since markers can propagate in parallel independently, SNAP exploits the maximum parallelism which the problem involves. Several AI applications such as natural language processing and classification system have been developed on SNAP.

WordNet developed at Princeton [4] is a computer based dictionary covering the vast majority of nouns, verbs, adjectives and adverbs from the English language. The words in WordNet are organized into synonym sets, which represent concepts. WordNet has a rich set of relation links among concepts. The relation links include `isa`, `part_of`, `causation`, `entailment` and others. Because of the rich set of relation links and its applicability to broad English, we decide to build our knowledge base on top of WordNet.

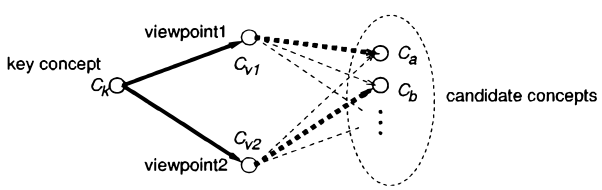


Fig. 1 Viewpoint-based similarity discernment problem.

Manuscript received December 17, 1997.

[†]The author is with NTT Communication Science Laboratories, Yokosuka-shi, 239–0847 Japan.

[††]The authors are with Faculty of Department of Computer Science and Engineering Southern Methodist University, Dallas, TX, U.S.A.

3. Viewpoint-based Similarity Discernment on SNAP

The algorithm for the viewpoint-based similarity discernment is presented below. It is based on the semantic connection between concepts and a viewpoint represented as one or more concepts in the knowledge base. The algorithm relies on marker-propagation. First, the construction of the knowledge base acquired from WordNet is explained, and then the algorithm is described in detail.

3.1 Construction of the Knowledge Base

WordNet is provided as a set of structured database files and search engine programs. A semantic network knowledge base can be build by a program that extracts information about concepts and their relation links from the structured files and converts them into a SNAP's representation scheme. It is also possible to acquire the sub-graph of the semantic network corresponding to a concept by following links connecting that concept with other nodes. While the sub-graph includes only a limited number of nodes, it is however a complete set of nodes related to a concept. Thus, an experiment and a simulation done on the sub-graph build in such a way provide almost the same results as in the case of a complete semantic network.

WordNet also includes glossary information for each concept, expressed in English. To be useful, this needs to be transformed into a semantic network form. We incorporate the defining features into the semantic networks by adding "`def_feat`" links between a concept and concepts included in that concept's glossary.

A portion of the semantic network for the concept `car#n1` is shown in Fig. 2 as an example. A circle corresponds to a concept node and an arrow line corresponds to a relation link. The name of a concept is comprised of the word which the concept corresponds to, part of speech and sense number for the word as they appear in the WordNet. For instance, `car#n1` represents that it is a noun concept and corresponds to the first sense for a noun word "car." Because SNAP provides directives for creating nodes and links, the mapping of the acquired semantic networks into SNAP is straightforward.

3.2 An Algorithm for Viewpoint-based Similarity Discernment

One way to measure the similarity between two concepts is to determine how many properties are shared among the two concepts. To be more precise, it depends on the ratio of shared properties to all of their properties. Viewpoint-based similarity can be defined

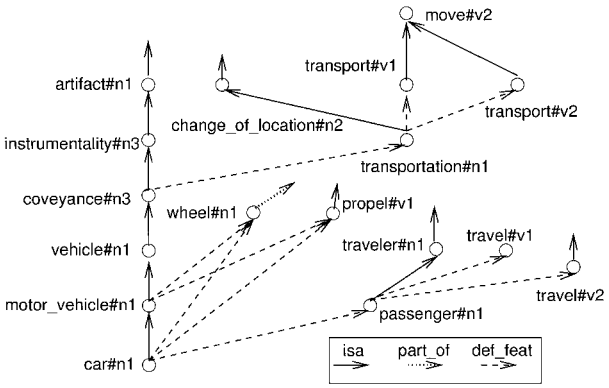
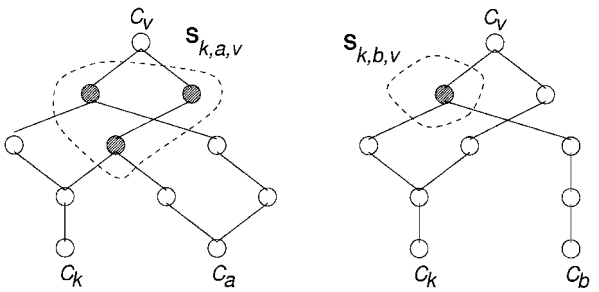


Fig. 2 An example of a semantic network acquired from WordNet.

in the same way as above, except that properties contributing to the similarity are restricted only to those having semantic connections with the viewpoint. In our knowledge base, concept properties are represented implicitly in the hierarchical structure. A chain of nodes connected with a concept node represents collectively the properties of that concept. When concepts are considered, the ratio of shared nodes to all nodes in the chains of nodes reflects the ratio of shared properties to all properties for the concepts. Therefore, comparing the ratio of shared nodes for two candidate concepts achieves the similarity discernment. Restricting the nodes involved in the comparison to those having semantic connection with a viewpoint achieves the viewpoint-based similarity discernment.

Based on the above discussion, we propose an algorithm for the viewpoint-based similarity discernment on SNAP. First, the algorithm tries to find semantic paths between a viewpoint concept on one hand and the key concept and the candidate concepts on the other hand. A chain of nodes and relation links that connect concepts C_x and C_y is called a semantic path between a concept C_x and a concept C_y. There may be several semantic paths between concepts. The paths between a concept and a viewpoint concept restricts the chains of nodes connected with the concept to those which are related to the viewpoint. We denote the set of nodes along the set of semantic paths between C_x and C_y as $\mathbf{P}_{x,y}$ and the set of shared nodes in the sets $\mathbf{P}_{x,z}$ and $\mathbf{P}_{y,z}$ as $\mathbf{S}_{x,y,z}$. Figure 3 illustrates examples of paths.

After finding the paths, the algorithm counts the number of nodes in $\mathbf{P}_{k,v}$, $\mathbf{P}_{a,v}$ and $\mathbf{P}_{b,v}$. Let them be $N_{k,v}$, $N_{a,v}$ and $N_{b,v}$ respectively. The number of nodes in sets $\mathbf{S}_{k,a,v}$ and $\mathbf{S}_{k,b,v}$ are also counted and represented as $M_{k,a,v}$ and $M_{k,b,v}$. Then, the ratios of shared nodes to all nodes for each candidate concept are computed. For concept C_a the ratio is $R_{k,a,v} = 2 \times M_{k,a,v}/(N_{k,v} + N_{a,v})$ and for concept C_b the ratio is $R_{k,b,v} = 2 \times M_{k,b,v}/(N_{k,v} + N_{b,v})$. Finally, the ratios $R_{k,a,v}$ and $R_{k,b,v}$ are compared and the concept which



(a) Paths $\mathbf{P}_{k,v}$ and $\mathbf{P}_{a,v}$ (b) Paths $\mathbf{P}_{k,v}$ and $\mathbf{P}_{b,v}$.

Fig. 3 Paths between the concepts and the viewpoint.

has larger ratio is chosen to be more similar to the key concept than the other concept.

The path finding procedure plays a significant role in the algorithm because it has to search the semantic connections between concepts among a vast number of node chains connected with each concept. The procedure to find paths between C_x and C_y on SNAP is briefly described as follows:

1. Set **Marker1** on C_x and **Marker2** on C_y.
2. Propagate the markers forward along any relation links.
3. Find the nodes where **Marker1** and **Marker2** collide and set **Marker3** on them.
4. Propagate **Marker3**s backward along any relation links.
5. Set **Marker4** on the nodes which have either both **Marker1** and **Marker3** or both **Marker2** and **Marker3**.
6. The nodes having **Marker4** belong to the path.

This procedure is targeted to SIMD-type SNAP (SNAP-1). In order to gain more flexibility and to allow for an asynchronous operation, a MIMD architecture and algorithm are more desirable. A MIMD-type SNAP (SNAP-2) has been developed and a path finding procedure for it has also been proposed [5].

4. Experimental Results

This section describes experimental results obtained with the algorithm for discerning between $SD\{$`travel#n1`, `horse#n1`, (`car#n1`, `ski#n1`)$\}$ and $SD\{$`sport#n1`, `horse#n1`, (`car#n1`, `ski#n1`)$\}$. A small scale knowledge base comprising of 260 nodes was constructed and the algorithm was executed on the SNAP-1 simulator.

Table 1 shows the nodes along the paths found by the path finding procedure.

Since $R_{\texttt{horse\#n1,car\#n1,travel\#n1}}$ is greater than $R_{\texttt{horse\#n1,ski\#n1,travel\#n1}}$, the algorithm concludes that `car#n1` is more similar to `horse#n1` than `ski#n1` under the viewpoint `travel#n1`. On the other hand, it concludes that `ski#n1` is more similar to `horse#n1` when `sport#n1` is given as viewpoint.

Table 1 Results of the path finding procedure.

g C_x	C_y	$\mathbf{P}_{x,y}$	$N_{x,y}$
travel#n1	horse#n1	riding#n2, ride#v1, ride#v2, travel#v5, travel#v1, transportation#n1, change_of_location#n2	7
travel#n1	car#n1	passenger#n1, conveyance#n3, motor_vehicle#n1,transportation#n1, wheeled_vehicle#n1, travel#v1, vehicle#n1, wheel#n1, change_of_location#n2	9
travel#n1	ski#n1		0
sport#n1	horse#n1	riding#n1, diversion#n1	2
sport#n1	car#n1		0
sport#n1	ski#n1	diversion#n1	1

5. Consideration and Future Work

We consider that similarity between concepts depends on concepts' shared properties, where properties are restricted only to those involved with a viewpoint. According to the experimental results, the path finding procedure, which is the key portion of the algorithm, was found feasible to restrict the concepts which reflect properties. Therefore, the algorithm is expected to achieve fairly accurate viewpoint-based similarity discernment. Accuracy evaluation comparing with human's discernment is future work.

The algorithm assumes that all properties involved with a concept are equally significant. However, in reality properties have different significance according to the relation connecting them to a concept. The accuracy of similarity discernment may be improved using marker values and link weights on SNAP that correspond to property significance. Accuracy improvement taking advantage of these features is also future work.

References

[1] K. Matsuzawa, T. Ishikawa, and T. Kawaoka "ABOUT Project for a robust problem solving and its method of measuring semantic similarity, " Japanese Society for AI, SIG-J-9401-14, 1994 (In Japanese).
[2] K. Kasahara, K. Matsuzawa, T. Ishikawa, and T. Kawaoka, "Viewpoint-based measurement of semantic similarity between words," Proc. Int. Workshop on Artificial Intelligence and Statistics, 1995.
[3] D. Moldovan, W. Lee, and C. Lin, "SNAP: A Maker-Propagation Architecture for Knowledge Processing," IEEE Trans. Parallel and Distributed Systems, vol.3, no.4, 1992.
[4] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Five papers on WordNet," Princeton Univ., CSL Report 43, 1990.
[5] S. Harabagiu and D. Moldovan, "A marker-propagation algorithm for text coherence," Parallel Processing in AI Workshop, IJCAI, 1995.