Learning the Space of Word Meanings for Information Retrieval Systems

Koichi HORI, Seinosuke TODA and Hisashi YASUNAGA

National Institute of Japanese Literature
1-16-10 Yutakacho Shinagawaku Tokyo 142 Japan

Abstract: Several methods to represent meanings of words have been proposed. However, they are not useful for information retrieval systems, because they cannot deal with the entities which cannot be universally represented by symbols.

In this paper, we propose a notion of semantic space. Semantic space is an Euclidean space where words and entities are put. A word is one point in the space. The meanings of the word are represented as the space configuration around the word. The entities that cannot be represented by symbols can be identified in the space by the location the entity should be settled in. We also give a learning mechanism for the space. We prove the effectiveness of the proposed method by an experiment on information retrieval for the study of Japanese literature.

1. Introduction

There have been no theories of semantics we can rely on for building a large information retrieval system. The defect in the existent theories is the lack of explanation of the mechanism for adjusting to the real world the formal symbolic systems used in the theories; the only thing they explain is the relation between natural language and the formal system. Those theories assume the existence of fixed and universal one-to-one relations between the basic elements in the formal system and the entities in the real world. For example, both Montague semantics and the situation semantics assume that we can represent the dog named Morris in the real world as some symbol like MORRIS in the formal system and that the relation between Morris and MORRIS is fixed and universal [3,2].

However, when we consider an information retrieval system, especially in the field of study on literature, we encounter problems where the assumption does not hold. One problem is that there are entities that do not have universal symbolic representation. For example, when a researcher discovers a new entity(or notion) in literature and writes a paper on that entity, the paper must be stored in the database but we do not have appropriate key words for that entity. When the entity becomes well known in later years, it may be named, for example, 'overthereism'. However at the time the entity is discovered and does not have the name 'overthereism', we must represent the entity by a fixed set of symbols, but it is not easy. Another problem is that the range of what is meant by a symbol differs among the users of an information retrieval system. For example, we cannot identify the fixed meaning of 'romanticism'. Every user assumes different meanings of 'romanticism' and it is not easy to control the meaning. The latter problem has been considered in the studies of fuzzy meanings, but, so far, the former problem has not been considered in the studies on semantics.

In order to solve the above mentioned problems, we propose a notion of semantic space and the learning mechanism of the space. Our assumption is that the entities which could not be represented by a fixed set of symbols can be identified in some semantic space by the location the entity should be settled in. Although whether this assumption is universally valid is problematic, we have proved that this assumption is effective in information retrieval systems in the field of studies on literature. We believe that the field of literature includes essential problems and has just enough complexity to give as evidence for a general discussion on semantics.

The semantic space is an Euclidean space where entities and words are scattered. The crucial point of our idea is that the axes of the space are not given beforehand but are generated through learning from

the interaction between a user and the information retrieval system. Since the axes of the space are not given beforehand, the system can adjust the configuration of the space for absorbing new entities.

In chapter 2, we describe in detail the notion of semantic space, explaining what are the entities and words in an information retrieval system for literature studies, and we show how the meanings of words are represented in the space.

In chapter 3, we describe the learning mechanism of the semantic space. Generally speaking, in the studies on machine learning, it has been revealed that the mechanism for controlling the learning process is important; without such mechanisms, the result of the learning becomes too general or too specific. In the learning process proposed in this paper, we use a user's satisfaction as the controlling criterion for learning. The result of the learning is a semantic space that just mirrors the world of literature existing in the user's mind. The reason we use the term 'learning' instead of 'acquiring' is that the information the system gets is not the direct expression of the meanings of words a user has in his mind but indirect and partial information given through the interaction between a user and the information retrieval system.

In chapter 4, we evaluate the effectiveness of the proposed ideas through an experiment. It is shown that entities that could not be retrieved by conventional key words can be retrieved in our system.

In chapter 5, we refer to related works and summarize our contribution.

2. Semantic Space

Before giving formal discussion, we first give an example of semantic space. Fig.1 shows an example of semantic space for philosophical issues in artificial intelligence(*).

We can find in the space shown in Fig.1, the book titled "Goedel, Escher, Bach" written by Hofstadter, as one entity. It has an internal ID to point to the information actually stored in bibliographical database. Since what is mentioned in the book has relation with Goedel, Escher and Bach, Goedel, Escher and Bach are located near the book. It must be noted that there has been no assurance, in previous theories, that Goedel denotes the famous mathematician everyone knows. However, in our semantic space, Goedel cannot be arbitrary things because the entity is bounded by other entities, some of which are pointed to actually stored information in database. The book "Goedel Escher Bach" describes some new notion. Since the notion was not known before the book was published, the notion does not have a universal name. We must read the whole thick book to know exactly what the notion is. In other words, symbols to represent the notion are equal to the whole book itself. However, we can determine the position of the notion, because we know the notion has relation with wholism, reductionism, mu, Goedel, Escher, Bach and so on.

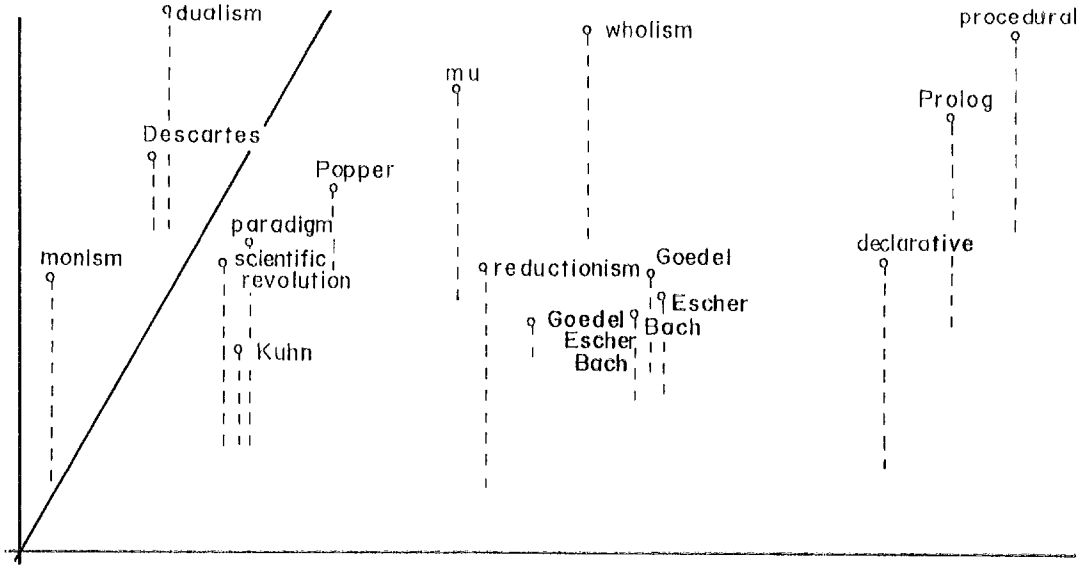

Fig.1 An example of semantic space

(*)Fig.1 is an explanatory example for the readers who are not familiar with Japanese literature. The semantic space made by our system is for Japanese literature and is authorized by the researchers in that field. The real example is given in chapter 4.

If the notion gets a name such as 'Hofstadterism' later, the name will be put in the position just above the book of Hofstadter. Until then, the notion is identified as some blank space above the book.

A user who wants to get a paper or a book on the same notion as one written in Hofstadter's book can find the notion in the space by looking at the configuration of the space.

One might think that we can represent the new notion intensionally. However, in practical information retrieval systems, it is difficult to fix the set of primitives for representing intension.

One might also think that identification of the location is similar to making conjunction of several keywords or to using a thesaurus. However, our semantic space has the prominent feature, which is absent in using keywords or thesaurus, that a user can embed his own ideas of meanings in the space. The only criterion for determining the space organization is a user's satisfaction. Every user has his own space in our system, and the system designer doesn't care whether one direction in the space denotes ISA relation or synonymic relation or else. Even the user himself may not know what kind of relation the axes represent in the space. However, as far as the user has his own idea about how one entity is related with other entities, the semantic space mirrors the world in the user's mind, and the axes (maybe in local subspace) as a result play the role of representing the relation between the entities such as ISA relation, synonymic relation, temporal relation, spatial relation, or a more complex relation.

So far, we have already described the main ideas of semantic space through an example. Now we give formal discussion.

The first question is what is the dimension of the space. The dimension of the semantic space is not given at first in principle. The dimension is determined as a result of space synthesis. Mathematical theories for calculating the dimension are given in theories of multi-dimensional synthesis developed in the studies of statistics. In our implementation, we limit the dimensions to three for the sake of simplicity of the system. Moreover, we give one fixed meaning to one dimension, that is, we give the meaning of symbolization to the vertical axis. At the bottom of the space, titles and authors of papers are arranged. At the top of the space, words which users use are arranged. At the middle of the space, writers and literary works and some controlled notions such as 'stylistics' are arranged. The reason we named the vertical axis 'symbolization' is that the upper space is a more symbolized world from the viewpoint of a bibliographical database.

The second question is what words and entities are. In our semantic space, we don't make clear distinction among the terms 'entity', 'notion' and 'word'. What exist in our space are, in any way, just symbols. But conventionally, we call the symbols that are used by users, 'words'; the ones that are pointed to bibliographical information 'entities(papers)', the ones that denote writers 'entities(writers)', the ones that denote works 'entities(works)', and others, 'notions'. As for the symbols that denote writers and works, they are controlled by an independently developed database for authentication.

The third question is what is meaning. We define the meaning of one symbol as the space configuration around the symbol. Moreover, we extend the notion of symbol to any point in semantic space, so that we can treat entities for which a symbol is not assigned. In other words, we identify a symbol with a point in space. For example, what the location for 'Hofstadterism' means is something between wholism and reductionism. You can understand what 'something between' in the previous sentence means by looking at the configuration of the neighbor space. For example, if you see neighbor 'Prolog' between 'declarative' and 'procedural', you know that 'something between' means the same kind of relation as 'Prolog' between 'declarative' and 'procedural'. It must be noted here that the same symbol can be put in more than one location. This allows a symbol to have several meanings depending on context.

The last question on the semantic space is what the distance is. The measure of distance exists in the user's mind. This is not elusion but essence. There is no explaining by what measures one researcher on literature judges that one entity is near another entity. It is because each researcher has his own measures that original papers can be written [personal discussion with some researchers on literature]. Since the semantic space is built for each user, the papers of the authors who have a quite different semantic world from that of the user may not have "proper" location in the semantic space. But that causes little problem, because such papers do not interest the user much.

3. Learning the Space

The outline for the use and learning of semantic space for an information retrieval system is as follows:

1. A user gives a query to the information retrieval system. The query is recognized as a sequence of words. Parsing of the query sentence is not done. So users generally give what they think are key words for search. For example, a user who wants a paper on the influence of Goethe on modern Japanese literature asks the system 'Goethe modern Japanese literature'.
2. The system searches in the semantic space for the same words the user gave. If such words are found, the system presents to the user the neighbor spaces of the words. If no such words are found, the system presents an

overview of the whole space (mainly the middle space, i.e. the space for writers and works).

3. The user selects some subspaces that attract him.
4. The system shows the details of the bottom spaces of selected subspaces.
5. The user selects papers from among those shown in the bottom spaces.
6. The system re-constructs the semantic space so that the selected papers in step 5 are located in shorter distance and selected subspaces in step 3 are located in shorter distance, and then the system puts the query words in the location above the selected papers.

We have implemented a system called ML0(Model Learner version0) that realizes the above mentioned steps. Fig.2 shows the configuration of the system. The system is written in Lisp and PL/I.

The monitor monitors all the functions of the system. It has special variables named *inconsistent and *attention.

*inconsistent is the variable for storing a pair of entities for which the distance in the semantic space is different from the estimated distance. The estimation of the distance is done as follows. When the initial semantic space is built, the distance between two papers is estimated, with some normalization, by the inverse of the number of occurrences of same words in the titles, and the distance between two

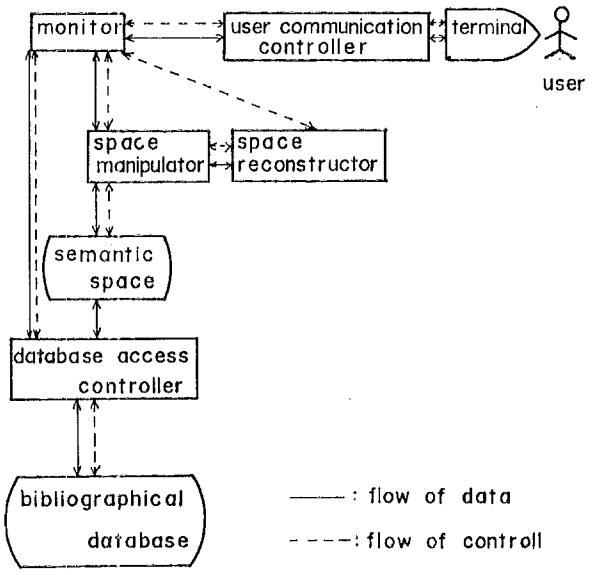

Fig.2 Configuration of the system

entities (other than papers) is estimated by the inverse of the number of the papers which include both entities in title. When the semantic space is reconstructed, the distance between entities which a user selected is estimated to some fixed small value, and the distance between entities which the system presented, but only one of which the user selected, is estimated to some fixed large value. The monitor judges that a user is satisfied if the real distance in semantic space is the same as the estimated distance. When the monitor detects the user's dissatisfaction, i.e. the difference between the real distance and the estimated distance, it registers in *inconsistent the pair of entities which caused the problem.

*attention is the variable for limiting the space for consideration. The monitor monitors the space only in the scope of *attention. This improves the efficiency of search and reconstruction.

The monitor triggers the space reconstructor after one session of query and answer if *inconsistent has value.

The space reconstructor plays the role of reconstructing the semantic space so that the user can be satisfied. It uses a heuristic procedure for space reconstruction mentioned below.

1. Select one pair from *inconsistent. (In the current version of the system, the pair which caused the largest inconsistency is selected.)
2. Inspect the density of the neighbor space for each entity in the pair, and decide to move the entity with less dense neighbors.
3. Enumerate the posiible new positions for the moving entity. (In the current version of the system, there are eight new candidate positions around another entity where the distance between the two entities is equal to the estimated value.)
4. Select from among them one position which causes the least new inconsistency.
5. Check new inconsistencies and register them in *inconsistent.
6. Go to step 1.

The monitor monitors the whole reconstruction process and stops the process by raising the threshold to judge the inconsistency when it judges that the reconstruction takes too much time.

Fig.3 shows an example of the process of space reconstruction. In Fig.3(a), the distance between the entities A and B was 10. Let's assume that a new estimation for the distance is 5. The reconstructor looks around the neighbors of both entities, and decides to move the entity B because the neighbors of B are less dense than those of A. The reconstructor selects one position that causes the least new inconsistency, from among eight positions around A, for B to be placed in. In Fig.3(b), B is placed to the left of A. New inconsistencies in the scope of *attention such as inconsistency about B and G are checked and registered in *inconsistent. After a few

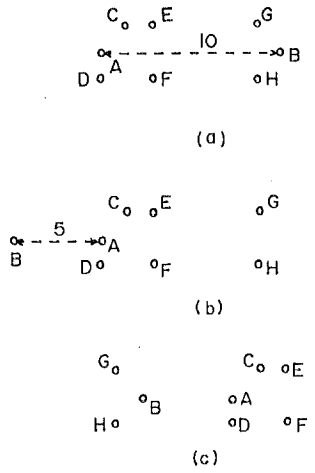

Fig.3 An example of space reconstruction process

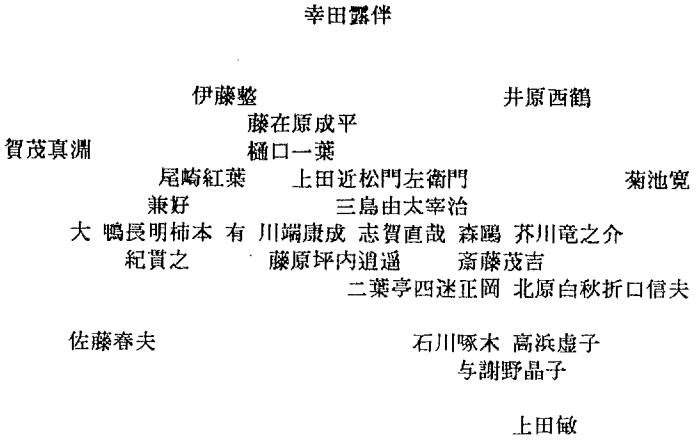

Fig.4 An example of initial semantic space (a horizontal section)

trial loops to decrease inconsistency, the space settles in the configuration shown in Fig.3(c), which includes no inconsistency.

Of course we can use more mathematical methods (e.g. matrix transformation of distance) for space reconstruction. However, the above mentioned heuristic procedure works more efficiently than mathematical methods, because so many pairs causing inconsistency are not detected at once due to the limitation of attention and rather small density of the world of literature.

4. Evaluation

First of all, we estimated the size of the semantic space. The number of writers studied in the field of research on Japanese literature is about 1900. The number of works studied is about 2300. Fortunately, these numbers are almost fixed. The number of papers written in that field in a year is about 5000. However, interest of one researcher is limited to less than one tenth of them.

We then confirmed, by asking some researchers and by analyzing the process of manual editing of research paper catalogues, that almost all notions in research on Japanese literature can be placed in positions among writers and works. For the notions that cannot be placed among writers and works, about 100 notions such as dialectology were prepared for building an initial space.

Finally we carried out an experiment by using the implemented system. Estimating the size of the semantic space for one user, we made an initial space in which 50 writers were registered. For building the initial space, we made a 50*50 distance matrix for the writers based on the information from the titles of 53563 papers written in the last ten years. The initial space made by the system is shown in Fig.4. The initial space itself is interesting enough for literature study. We can know from the space the striking fact that the poet named Bashou in the seventeenth century has a strong relation with many modern novelists. We can also infer that many researchers have special interest in female writers, finding the subspace where ancient and modern female writers are clustered.

We asked a researcher on Japanese literature to search papers based on a complex query. One query was "on the process of transformation of fables into written forms, how fables were transferred among people, ecology of fables"(*). Looking at the semantic space, he first roughly enclosed in the space about ten works which he thought to be related to "fables". Glancing at the bottom space of them, where 1337 papers (written in the last ten years) were scattered, he could point out 112 papers which matched the query.

The selected papers which included in the titles words such as nature, development, transfer, fablization, or generation were gathered into one subspace by the system after the session, and the user's query was registered above the subspace. This means that, next time, the system can give the answers to the

(*)The query was given in Japanese. This is a translation made by the authors.

same query at once, and can register, in the proper subspace, new papers on the same notion.

The greatest merit of the semantic space was that the system could tell the user what it knew about the things related to the user's request. In conventional systems which depend on key words, users must imagine what words the system knows; in fact, the subject of the experiment could not give proper key words for the complex query. In contrast with this, the ability of the user to point out the subspaces he wants by just glancing at the whole presented space was more than we expected.

5. Related Works

The most related work is Rieger's[7,8]. He also made some kind of semantic space from the information of relations among words. However, he made the space based only on the frequency of co-occurrences of words in sentences. Our semantic space can have more reality about meanings than his, because symbols that are actually connected to entities stored in the database exist in our space.

In the sense of treating entities which do not have symbolic representation, connectionist models[4,9] have some relation with our approach. However, it seems that the connectionists have not yet realized the process of yielding symbols on connection networks that can represent such abstract notions as dealt with by our system.

From the methodological point of view, methods developed in the studies of multi-dimensional synthesis in statistics are related to ours. The reason we have developed methods to calculate the space independently from those methods is similar to one mentioned by Lebowitz[6]. That is, we are not interested in mathematical rigidness of the process of analyzing data, but in modeling what occurs in the human mind.

From the viewpoint of information retrieval systems, there are many more requests than those answered by our approach. They include treating structures of key words more explicitly, understanding the user's intention from his query[1], or making co-operative responses[5]. However, the previous studies on those problems started ignoring the most basic problem of understanding and learning the meanings of one word. Our work can give an assured starting point for those further studies.

6. Conclusion

We proposed to represent the meanings of words in space. We gave a learning mechanism for the space. We proved the effectiveness of the proposed methods in an information retrieval system.

One of the reasons we chose the field of Japanese literature was that there was strong demand from the researchers in that field to make a useful system; systems based on conventional key words did not work well for the field of literature.

Another reason is that the field offers good examples of entities for which universal symbolic representation is difficult.

To apply our method to other fields, we must solve only two problems. One problem is to select symbols to be put in the initial space. In literature, works and writers played the role. The other problem is to determine criteria for space configuration. We used a user's satisfaction with the answers from the information retrieval system as the criterion. If these two problems are solved, our method can be applied to any domain. These problems do not seem so difficult.

From the viewpoint of artificial intelligence, the semantic space gives a basis for studies on abduction and analogy. The discovery of blank space surrounded by symbols can lead to discovering of new ideas by machines. Since we can measure semantic similarity directly by distance in semantic space, we can make analogical reasoning based on that similarity. For the same reason, from the linguistic point of view, the semantic space can be useful for understanding metaphorical expression such as "Hofstadter is Prolog between procedural semantics and declarative semantics".

References

[1]Allen,J.:Recognizing Intentions from Natural Language Utterances, in Brady,M.,Berwick,R.C.(eds.): *Computational Models of Discourse*, MIT Press(1983).

[2]Barwize,J.,Perry,J.: *Situations and Attitudes*, MIT Press(1983).

[3]Dowty,D.R. et al.: *Introduction to Montague Semantics*, D.Ridel Publishing(1981).

[4]Feldman,J.A.:Connectionist Models and Their Applications Introduction, *Cognitive Science*, Vol.9,pp.1-3(1985).

[5]Kaplan,J.:Cooperative Responses from a Portable Natural Language Database Qury System, in Brady,M.,Berwick,R.C.(eds): *Computational Models of Discourse*, MIT Press(1983).

[6]Lebowitz,M.:Categorizing Numeric Information for Generalization, *Cognitive Science*, Vol.9,pp.285-308(1985).

[7]Rieger,B.B.:Procedural Meaning Representation by Connotative Dependency Structures. An Empirical Approach to Word Semantics for Analogical Inferencing, *Proc. COLING82* (1982).

[8]Rieger,B.B.:Semantic Relevance and Aspect Dependency in a Given Subject Domain; Contents-Driven Algorithmic Processing of Fuzzy Wordmeanings to Form Dynamic Stereotype Representations, *Proc. COLING84* (1984).

[9]Rumelhart,D.E.:Feature Discovery by Competitive Learning, *Cognitive Science*, Vol.9,pp.75-112(1985).