Converting Language Computation into Mathematical Operations

Chengming Guo
Computer Science Department
Tsinghua University
Beijing 100084 China
E-mail: chengming@bepc2.ihep.ac.cn

ABSTRACT

Recent development in machine-readable dictionary (MRD) research at Tsinghua University makes it possible to turn lexical data into unique mathematical representations. The mathematical representation can be converted back to symbolic data, when necessary, without losing any original lexical information. Mini-LDOCE, a dictionary of the definition primitives of Longman Dictionary of Contempoary English (LDOCE) is being used as the testing ground for the bijectional technology. Research on lexical data bijection represents part of the effort to reduce language computation to mathematical operations, the focus of research being the establishment of correlates of vector/matrix computations with the computation on language meaning.

Lexical data is being converted into unique mathematical representations by a process called "lexical data bijection", which refers to the bi-directional mapping from lexical data in symbolic form to vector representations, and then back from vector representations to the original symbolic form of the lexical data. Research on lexical data bijection represents part of the effort to reduce language computation to mathematical operations, the focus of research being the establishment of correlates of vector/matrix computations with the computation on language meaning.

One important development in Artificial Intelligence (AI) and Cognitive Science research in recent years warrants the attention of lexical researchers. It involves the trend for the integration of natural language processing (NLP) with various subareas of AI, e.g., computer vision [1,2]. The need for integrated representation schemes incorporating both perceptual information and common sence knowledge poses new challenges.

Recent development in machine-readable dictionary (MRD) research at Tsinghua University makes it possible to turn lexical data into unique mathematical representations. The mathematical representation can be converted back to symbolic data, when necessary, without losing any original lexical information. The mathematical representation of the encoded lexical data takes the form of multi-dimensional vectors. Mini-LDOCE, a dictionary of the definition primitives of Longman Dictionary of Contempoary English (LDOCE) is being used as the testing ground for the bijectional technology.

This paper attempts to give an account of our current understanding of the issue.

1. Descriptive vs Prescriptive Semantic Primitives

Essential to the learning of new concepts, new rules, and new domain theories by machine is the existence of a set of hierarchically arranged primitives [3,4]. Such primitives take the form of semantic primitives [5,6] in the present study. Two alternative approaches to the development of a set of semantic primitives exist — the *prescriptive* approach and the *descriptive* approach. In the prescriptive approach, a set of primitives is defined, or *prescribed*, prior to, or in the course of designing and developing a system. An example of a prescribed set of semantic primitives is the set of semantic features used as *box* codes in the electronic version of LDOCE. The descriptive approach [6, p. 198], on the other hand, allows a natural set of semantic primitives to be derived from a natural source of data such as a dictionary. The set of definition primitives derived from a particular MRD closes on the MRD. A descriptive set of semantic primitives of a particular natural language closes on that language. It is believed that although it is feasible to derive a set of definition primitives from one particular

MRD, a descriptive set of semantic primitives, true of one natural language, is preferably derived from more than one natural source such as a dictionary.

Although the theoretical implications and computational consequences of a natural set of semantic primitives are difficult to ascertain at this point, these primitives can be compared with the formalized language of pure thought which Frege called the *Begriffsschrift* [7]. Frege's system was modeled upon the language of arithmetic, such that every thought is composed of the primitive ideas represented by the primitive signs of the *Begriffsschrift*. The natural set of semantic primitives derived from natural sources is itself part of the natural language people use to communicate thought. The fundamental theoretical assumptions underlying the derivation of a natural set of semantic primitives from machine readable sources share much in common with the theories of cognitive grammar, the earlier version of which was Space Grammar [8], namely:

1. Although any interestingly strong version of the Whorfian hypothesis, which essentially claims that language determines thought, is dubious, it is undeniably true that languages embody divergent codifications of conceived reality.

2. The putative difference between linguistic and extralinguistic knowledge, or between a dictionary-type account of the meaning of lexical items and an essentially encyclopedic account, is illusory.

Point 1 above necessitates the derivation of different sets of semantic primitives for different languages. A natural consequence of 2 concerns an understanding that language and knowledge are ultimately inseparable.

2. LDOCE and Mini-LDOCE

LDOCE is a full-sized dictionary designed for learners of English as a second language containing over 55,000 entries in normal book form and 41,100 entries in machine-readable form (a typesetting tape). An *entry* is a collection of one or more sense definitions that ends at the next head. The *head* is the word, phrase or hyphenated word defined by an entry. A single word can be the head of more than one entry if homographs or different parts of speech exist for that word. A *sense definition* is a set of definitions,

examples and other text associated with one sense of a head. If an entry includes more than one sense definition, then each sense definition will have a number.

Table 1 below shows some basic data derived from Plate's analysis [9] of the machine-readable tape of LDOCE (because of a tape error, words that follow alphabetically after "zone" have not been analyzed). The figure of a 2,166 word controlled vocabulary is arrived at as follows. The list of controlled vocabulary given in LDOCE contains 2,219 items. Among them 58 are prefixes and suffixes. Thirty-five other words from the list did not have heads. These are all removed. Furthermore, the analysis shows that some words are not part of the controlled vocabulary yet are used frequently in definitions, for example, the word "aircraft" is not part of the controlled vocabulary yet it is used 267 times in sense definitions. About thirty such words have been added to the list of controlled vocabulary words, giving a total of 2,166.

Heads	Words	Entries	Sense Definitions
Controlled Vocabulary	2,166	8,413	24,115
Non-Controlled Vocabulary	25,592	32,687	49,998
Totals	27,758	41,100	74,113

Table 1. Head Counts for words, entries, and senses in LDOCE.

The machine-readable version of LDOCE contains box and subject codes that are not found in the book. The box codes use a special set of prescribed primitives such as "abstract," "concrete," and "animate," organized into a type hierarchy. The primitives are used to assign type restrictions on nouns and adjectives, and type restrictions on the arguments of verbs. The subject codes use another special set of primitives organized into a hierarchy. This hierarchy consists of main headings such as "engineering" and subheadings like "electrical." The primitives are used to classify words by their subject, for example, one sense of

"current" is classified as "geography: geology" while another sense is marked "engineering/electrical."

It was found that a subset of the 24,115 controlled vocabulary senses, about 4,000 of them, are used to define the controlled vocabulary words themselves [10]. Mini-LDOCE is a dictionary of these 4,000 definition primitives of LDOCE. In other words, Mini-LDOCE has about 4,000 entries, each entry being a definition primitive of LDOCE. Mini-LDOCE represents continued work from Guo (1989) at New Mexico State University. The dictionary has been compiled by Guo for publication by Longmans.

Mini-LDOCE differs from LDOCE in the form of the word sense definition text. Whereas LDOCE word senses are defined by words, Mini-LDOCE definitions are given in word senses, i.e., each word in the definition text is already disambiguated, and has a sense number attatched at the end. The set of the controlled vocabulary word senses used in the definition text of Mini-LDOCE falls within the set of total Mini-LDOCE sense entries. When non-controlled vocabulary words, or phrases formed either of controlled vocabulary words, non-controlled vocabulary words, or a combination of controlled vocabulary and non-controlled vocabulary words are found in Mini-LDOCE sense definitions, an effort is made to disambiguate the non-controlled vocabulary word or phrase, and reduce its definition as found in LDOCE to the definition primitives of LDOCE. This results in embedded definition, which sometimes runs down three or four levels. In most cases, the embedded definition bottoms out to definition primitives within three levels of embedded definitions.

3. Lexical Data Bijection

Reported below is an account of the most recent developments in the work of the NLP group of China National Laboratory of AI Technology and Systems.

3.1. Guo's hypothesis

Over the years, one question has been haunting, i.e., given the type of sense definitions found in Mini-LDOCE, is there a function such that its operation over the sense definitions given in terms of *prescritive semantic primtives* can generate unique mathematical representations so that

its mapping back to the original sense definitions retains the original lexical information?

3.2. Chen's bijectional algorithm

Zushun Chen, member of our NLP group, is a mathematician by training. His bijectional algorithm has been presented to, and scrutinized by, well-known Tsinghua University mathematician Prof. Zhenhua Ma.

3.2.1. Background

The algorithm is based on a simple theorem of continued fractions in number theory, i.e., for any pair of mutual prime numbers (P,Q), there exists one *unique* simple continued fraction [A1,A2,...,An], where P, Q, and A1, A2, through An are natural numbers, and An is greater than 1.

3.2.2. Encoding algorithm in pseudo codes

The term *encoding* here refers to the mapping from lexical data to vector representations. Input data A1,A2,...,An are assumed to have already been put into the array IN, where the lenghth of n is known.

```
ENCODING(IN, n, P, Q)
      integer array IN;
      integer n, P, Q;
   Begin
      integer x1, y1, x2, y2, x3, y3, i;
      IN[n] = IN[n] + 1;
      x2=0; y2=1;
      x3=1; y3=0;
      for i =1 to n step 1
      begin
         x1=x2; y1=y2;
         x2=x3; y2=y3;
         x3=IN[i] * x2 + x1;
         y3=IN[i] * y2 + y1;
       end_of_for_i;
     P=x3; Q=y3;
End_of_ENCODING
```

3.2.3. Decoding algorithm in pseudo codes

The term *decoding* here refers to the mapping from vector representations back to the original symbolic data. The input data P,Q is assumed to be known, and the output data is assumed to have been put into the array OUT, whose lenghth is recorded in N.

```
DECODING(P, Q, OUT, N)
    integer array OUT;
    integer P, Q, N;
    value P, Q, N;
  Begin
    integer m, i, x, y;
    i=0; m=Q; x=P;
    while m != 0 do
     begin
       y=x; x=m;
       i=i+1;
       OUT[i]=[y/x]; /* integral part --incomplete quotien */
       m=y MOD x;   /* the least residue to modulo x */
      end_of_while_m;
    N=i;
    OUT[N]=OUT[N]-1;
End_of_DECODING;
```

3.3. Gong's implementation

Notice that the encoding process produces a number pair *P* and *Q*. They are converted into one number *F*, using the following scheme, i.e., the number of digits of *P* is coded into the first two leading bytes in *F*, followed by the number *P* and *Q*, with *Q* trailing *P*.

References

[1] Dennett, D. (1991). Consciousness explained. Harmondsworth: Penguin.

[2] McKevitt, P. (1994). (Guest Editor) Integration of natural language and vision processing. Special volume (Issues 1, 2, 3) of AI Review Journal Dordrecht: Kluwer.

[3] Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), The

psychology of computer vision . New York: McGraw-Hill.

[4] Dietterich, T. G., & Michalski, R. (1981). Inductive learning of structural descriptions. Artificial Intelligence , 16, 257-294.

[5] Wilks, Y. (1972). Grammar, meaning, and machine analysis of language. London: Routledge.

[6] Wilks, Y. (1977). Good and bad arguments about semantic primitives. Communication and Cognition, 10, 182-221.

[7] Frege, G. (1960). Translations by Geach, P. & Black, M. Blackwell: Oxford.

[8] Langacker, R. H. (1982). Space grammar, analyzability, and the English passive. Language, 58, 22-80.

[9] Wilks, Y. A., Fass, D. C., Guo, C-M., McDonald, J. E., Plate, T., and Brian, B. M. (1990). Providing machine-tractable dictionary tools. *Machine Translation.* 5, pp. 99-154.

[10] Guo, C-M. (1989) Constructing a Machine-Tractable Dictionary from Longman Dictionary of Contemporary English. Doctoral dissertation. New Mexico State University.