INTERACTIVE VOCABULARY ACQUISITION IN XTRA

Cheng-ming Guo
Computing Research Laboratory, Dept. 3CRL
New Mexico State University
Box 30001, Las Cruces, NM 88003-0001, USA
(505) 646-5466
CSNET: guo@ nmsu

ABSTRACT

This paper describes a practical solution to on-line dictionary update in XTRA, a machine translation system developed by Xiuming Huang at the Computing Research Laboratory of New Mexico State University. The focus of the discussion in on IVES — an Interactive Vocabulary Enrichment System built by this writer for XTRA, It reflects an ongoing effort at the laboratory to build embedded learning mechanisms in machine translation systems. Two types of learning are discussed, word learning and word sense learning. Each type of learning undergoes three routine processes: detection, acquisition, and evaluation. The emphasis of this paper is on the use of semantic preference violations in the detection of the need to learn new word senses.

I INTRODUCTION

Machine translation systems, built using an AI approach, could be equipped with embedded learning mechanisms so that the systems would constantly update themselves and modify their behavior. Building such embedded systems should speed up the early realization of machine translation on microcomputers. This is because, except for the closed system of function words, an MT procram cannot be expected to have a 'complete' vocabulary at the time of installation. And it is almost impossible for the customer to purchase such systems that are adequate to start with. Learning mechanisms could be a remedy in this regard. It is usually easier to find a bilingual person than a software engineer to upgrade a system.

The focus of the discussion is on IVES — an Interactive Vocabulary Enrichment System built by this writer for XTRA. Two types of learning are discussed, word learning and word sense learning. Each type of learning undergoes three routine processes: detection, acquisition, and evaluation. The emphasis of this paper is on the use of semantic preference violations in the detection of the need to learn new word senses.

II WORD LEARNING

According to a model of machine learning advocated by Cohen and Feigenbaum (1982), machine learning has four important elements: (a) the environment, (b) the learning element, (c) the knowledge base, and (d) the performance element. The environment supplies information to the learning element; the learning element uses this information to make improvements in the knowledge base; the performance element uses the acquired informa-Uon to carry out its task; finally, feedback information gained during attempts to perform the task goes back to the learning element. In the context of word and word sense leaming in XTRA, the environment is the system's interaction with the user; the learning element is IVES; the knowledge base is the system dictionaries; and the performance element is the English sentence parser. Depending on the kind of information supplied by the environment to the learning element, there exist five different learning strategies. These are rote learning, learning by instruction, learning by deduction, learning by analogy, and learning by induction. (Michalski, Carbonell, and Mitchell, 1986, p. 14).

Word learning in XTRA belongs to rote learning. In rote learning the information supplied by the environment is directly accepted by the learning system. The strategy is elementary, not powerful enough to accomplish intelligent learning on its own, but it is an 'inherent and important part of any learning system' (Cohen and Feigenbaum, 1982, p. 335).

The detection of the need to learn new words in XTRA is facilitated by XTRA's preprocessing routine. Whenever a certain word in the input sentence is found to be missing from XTRA's dictionaries, the preprocessing routine issues a warning and prints out the missing word. Whenever a word is detected as missing, IVES is called, and an interactive session with the system user begins.

During the acquisition phase, information is elicited from the user to expand XTRA's dictionaries. IVES, as it stands now, handles four major categories of open system words, i.e., nouns, verbs, adjectives, and adverbs. All input into the system dictionaries is checked for its validity. Invalid entries are automatically rejected, and a second request for the same piece of information follows the rejection. Although the default value of the target language is Chinese, IVES could handle any target language.

When the acquisition phase is over, parsing continues from where it left off. Often, it succeeds as a result of dictionary expansion. However, it may also fail for reasons that will be discussed in the next section.

III WORD SENSE LEARNING

Word sense learning can be categorized as a special type of learning by instruction. It undergoes the same three processes as word learning does, i.e., detection, acquisition, and evaluation. The difference between word learning and word sense learning is that the detection process for the latter is much more complicated than that for the former. Also, in the course of the detection process,

responses from the user are not learned by rote and used by the system. Instead, they are taken as pragmatic instructions and carried out accordingly. Word sense learning uses the same learning module as word learning to acquire new word senses. The acquired word senses are always evaluated by XTRA.

The need to learn new word senses in XTRA is detected when a semantic preference violation occurs. In semantic primitive based systems, semantic primitives and semantic preference rules embody world knowledge, and they play an extensive role in word sense disambiguations (Wilks, 1972; Huang, 1985). In fact, the same mechanism can be employed in detecting the need to learn new word senses. Suppose XTRA has the word 'bridge' in the svstem which means 'a physical structure built on the river', and it also knows the word 'play' as in 'play basketball'. When processing the input sentence 'John plays bridge', the system would fail as a result of semantic preference violations. In this case, the semantic specification of the expected object of the predicate verb (play) is 'game' whereas the semantic specification of the actual object head noun (bridge] is 'grain', which means 'any kind of structure'. What nap pens is that 'grain' is not anywhere close to the semantic specifications of the expected object head noun of 'play', hence a semantic preference violation. This particular type of semantic preference violation indicates the lack of a required word sense, and therefore the need to learn.

Unfortunately, not all violations of semantic preference rules signal the need to learn. Only some of the preference violations are learning signals. The non-learning signals are symptoms of some linguistic complications, such as ill-formedness. In word and word sense learning, ill-formedness often involves misspelling and the figurative use of language. It is not surprising that part of the game of word and word sense learning is the recognition and handling of metaphors and misspellings. IVES will not request new vocabulary information until the last recorded semantic preference violation is found unaccounted for: i.e., when no metaphors or misspellings are identified in any pair of recorded worcf senses involved in semantic preference violations.

A. Recognition and Handling of Metaphors

Many solutions have been suggested to handle ill-formeaness incurred by metaphors. Four strategies were proposed by Fass and Wilks (1983) for semantic primitive based systems. They are the passive relaxation strategy, the Change The Data (CTD) strategy, the Change The Expectations (CTE) strategy, and the active strategy that involves the use of pseudo-texts (Wilks, 1978). The strategy adopted by XTRA is to relax the preference of the predicate and accept the semantic representation with the conflict unresolved. Relaxation is a useful heuristic to handle metaphors, but not without flaws. To illustrate the unpredictable nature of the relaxation strategy, look at the following sentence:

The chopper drank gasoline.

'Chopper' can either be a 'helicopter' or an 'ax'. In this particular sentence, more than likely it means a 'helicopter'. What will happen if the system dictionary has (a) the 'helicopter' sense of chopper only, (b) the 'ax' sense only, or (c) both the helicopter' and the 'ax' senses? Since XTRA relaxes the preference of 'drank' for an animate subject in all three cases, we obtain the desired sense of chopper under (a); the less than desired sense ol chopper under (b); and unpredictable results under (c). In other words, 'relaxation' can help produce the correct parse sometimes, but may not do so at all times.

As Carter (1984) pointed out, to recognise metaphors 'recourse to a richer source of knowledge like Wilks' pseudo-text is necessary.' A pseudo-text is a preference semantic representation of factual and functional information about a concept. The form of representation used in IVES incorporates the pseudo-text idea in a form that integrates linguistic knowledge with general world knowledge in one unit of representation. It is called the Integrated Semantic Unit (ISU). Each ISU is an integrated representation of the meaning of a word sense. The building blocks of an ISU are also word senses. The general form of an ISU is as follows:

```
isu(Wordsense,
   belong( [Class]),
   ik(integrated knowl edge)
```

where 'isu' stands for integrated semantic unit; 'Wordsense' is a word sense of any entry word in *Longman Dictionary of Contemporary English* (LDOCE); 'belong' introduces a hierarchical relationship between Wordsense' and 'Class'; 'Class' represents a superordinate word sense in the hierarchy of word senses in LDOCE; 'ik' introduces integrated linguistic and world knowledge associated with Wordsense. Although the general form of an ISU is uniform across all four open-class categories of English words, i.e., nouns, verbs, adjectives and adverbs, the actual specification of their respective superordinate word senses, and integrated knowledge varies from category to category. An example is given below:

```
isu(chopper1,
   belong([tool1]),
   ik(part([handle1, blade1]),
      property([[cut1, meat1]]),
      part_of_speech([n]),
      root_form([chopper]),
      inflectional_form([choppers]),
      feature([]),
      co_existing_senses([3]),
      sem_head([thing]),
      p_phrase([])
      )
   ).
```

IVES uses ISUs to recognize metaphors. In fact the recognition process begins with the recording of all semantic preference violations detected during parsing. Some of these violations represent a searching process and are eliminated. Only interesting semantic clashes remain. IVES looks into each one of these clashes for indications of metaphorical use and misspelling before it requests to acquire any new word senses. Among the semantic clashes that IVES can handle are subject/verb mismatches, verb/object mismatches, adjective/noun mismatches, and subject/predicative adjective mismatches.

To recognize metaphors. IVES first tries to identify three important parts of a metaphor, the tenor, the vehicle, (Richards, 1936) and the salence (Ortony, 1970). Ortony takes 'salence' to mean an estimation of 'prominence of a particular attribute with respect to a concept to which it does or could apply.' (p. 162.) In 'The chopper drank gasoline', the tenor is man, a human-being who can drink liquid; the vehicle is 'chopper' the actual subject of the sentence; the salence is found in the resemblance between a man's act of drinking and whatever a chopper can do. Computationally, the expected subject of 'drank' is taken as the tenor; the actual subject 'chopper' is taken as the vehicle; and 'drank', the main verb of the sentence, is used to find the salence. IVES examines the properties associated with both the tenor and the vehicle to see if their properties are both related to the act of drinking. Once a metaphor is found, IVES would check with the user for confirmation. To handle metaphors, the CTE strategy as mentioned earlier is used. If no metaphors are spotted, IVES would start looking for spelling errors. The recognition and handling of misspellings are discussed in the next section.

B. Recognition and Handling of Misspellings

IVES recognizes two types of spelling errors in collaboration with the user. They are context--independent errors and contextrdependent errors. Context-independent errors can be spotted without examining the context where the error occurs. These spelling errors typically involve missing, surplus, or wrong letters, or letters with their proper positions switched, 'convenence', 'conveenience' conviniece', and 'conveneince' for 'convenience would be examples in point. These errors would be picked up by XTRA's preprocessing routine simply Because the misspelled forms are not found in the system dictionaries. The system would treat them as missing words and call the learning module to input the required information. To recognize and handle contextrindependent errors, the learning module checks on the spelling of any word that calls for new vocabulary information. The user is requested to do the actual checking, and is given a chance to correct him/herself when a spelling error is spotted. However, it so happens tnat sometimes the misspelled word takes the form of the correct spelling of another word which is already available in XTRA's dictionaries. We call this second type of misspellings contextrdependent errors. Their recognition requires the examination of the context where the error occurs. More often than not, the spelling error would cause a semantic preference violation, and IVES would undergo the same processes as described in subsection A above to process the semantic mismatch. When IVES fails to identify any metaphors, it will start looking for spelling errors. Again the user is requested to do the actual checking, and is given a chance to correct him/herself when a spelling error is spotted. Two research questions remain open and unanswered at this stage:

1. What if the user is locked in an undesireable mind set and fails to see the spelling error when he or she is given a chance to?
2. What if a context-dependent spelling error does not cause any semantic preference violations?

IV CONCLUSION

On-line vocabulary acquisition is not a recent phenomena either for machine translation in particular or for natural language processing in general. What is new in this paper is the use of semantic primitives and semantic preference violations in the detection of the need to learn new word senses. The mechanism originally designed to disambiguate multiple word senses in an MT system is taken advantage of and made into part of a complex of mechanisms intended to detect the lack of required word senses, and therefore, the need to learn. To cope with tne inescapable problem of metaphor recognition in the process of detecting the need to learn new word senses, a new form of representation, ISU. that integrates linguistic knowledge with general world knowledge is introduced. The motivation behind such introduction stems from the following observation:

Knowledge-based systems typically ignore the rich linguistic knowledge found in conventional dictionaries while conventional dictionaries never express the rich commonsense knowledge assumed of the reader explicitly. Yet, linguistic knowledge and general world knowledge are not ultimately separable. What natural language processing really needs is an integrated form of representation that combines the two kinds of knowledge in one unit of representation (Wilks, 1978). The present research represents an effort in the application of the integrated approach to the area or dictionary update in machine translation.

ACKNOWLEDGEMENTS

My deep appreciation is due to all members of our natural language processing group, particularly to Xiuming Huang, Dan Fass, ana Yorick Wilks, for their encouragement, advice and help.

REFERENCES

Carter, D. M. "On the Fass and Wilks proposal to use 'polysemy rules'." *Computational Linguistics* 10:2(1984) 147.

Cohen, P. R. and E. A. Feiganbaum *The handbook of artificial intelligence.* Stanford, Heuristech & Los Angels, William Kaufmann, 1982.

Fass, D., and Y. Wilks, "Preference Semantics, 111-formedness, and Metaphor." *Computational Linguistics* 9: 3-4 (1983) 178-187.

Huang, X-M. "Machine Translation in the SDCG Formalism." In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.* Colgate University, New York, 1985.

Michalski, R. S., J. G. Carbonell, and T. M. Mitchell (Eds.). *Machine Learning: an Artificial Intelligence Approach Volume 2.* Los Altos, Calif., Morgan Kaufmann, 1986.

Ortony, A. "The Role of Similarity in Similes and Metaphors." In: Ortony, A., (Ed.) *Metaphor and Thought.* New York, Cambridge University Press, 1979.

Richards. I. A. *The Philosophy of Rhetoric.* New York, Oxford University Press, 1936.

Wilks, Y. *Grammar, Meaning and the Machine Analysis of Language.* Routledge, London, and Boston, 1972.

Wilks, Y. "Making Preference More Active." *Artificial Intelligence* 10:1(1978) 1-11.