

Case Revisited: In the Shadow of Automatic Processing of Machine-Readable Dictionaries

Fuliang Weng

Computing Research Lab, New Mexico State University
Las Cruces, NM 88003

This paper discusses the work of automatically extracting Case Frames from Machine-Readable Dictionaries based on a three layer *a posteriori* Case Theory[5].

The theory is intended to deal with two problems:

1. To dynamically adjust grains of Cases. This is where *a posteriori* comes from.

2. To provide a procedure to determine Cases. This is where *three layer* comes from.

The three layers are:

1. **base layer:** This layer is intended to accomplish transformations of words to concepts by explicating language and word specific implicits. e.g., for the verb *eat* in the intransitive case, its subject is *eater*, while for verb *break* in the intransitive case, its subject is *the broken*.

2. **default layer:** in this layer, implicit assumptions of naive theories are made explicit, e.g., for concept *see*, there are two different views towards its subject: if a person who uses this concept believes that *seeing* is just a process of passive perception, then this person will assign to its subject, a *passive*¹ Case such as

experiencer; if a person who uses this concept believes that *seeing* is a process of active selection, then this person will assign to its subject, an *active* Case such as *agent*.

3. **context layer:** in this layer, Cases are further clarified upon any requests from current tasks, associated context and personal belief systems (knowledge), e.g., in sentence *The commander forced the soldier to break the door.*, whether the *soldier* should be assigned *agent*, *instrument*, *active*, or something else, should be decided by both contextual information and needs.

Arguments for the three layer theory can be found in[5].

Relevant knowledge sources for arriving at different layers are:

1. **Formation of the base layer:** the formation is based on knowledge sources which mainly come from syntactic codes and definitions in LDOCE (Longman Dictionary of Contemporary English). Examples in LDOCE also contribute to this process [1].

2. **Formation of the default layer:** the formation is based on the assumption that naive theories are weakly consistent, which implies that certain semantic classifications may be consistent with certain naive theories: verb, noun, preposition and adjective classifications based on semantic and pragmatic codes in LDOCE, and examples in LDOCE can help to obtain such theories.

3. **Formation of the context layer:** the unification of the base layer and the de-

*I would like to express thanks to Dr. L. Guthrie, Dr. D. Farwell and Prof. Y. Wilks for comments and encouragement. This project is supported in part by CRL. Some of the ideas were developed during my stay in CS/Fudan and CMT/CMU.

¹The words *passive/active* are used to indicate different levels of activeness. In what follows, Cases such as *agent* and *instrument* have somewhat different meanings than the conventional ones. We use them just for referring to a group of phenomena which are related to their names.

fault layer forms an initial representation of the context layer, its further development mainly depends on task, contextual needs and personal belief systems. The initial representation is a tuple with three components: entity-role, environment and endurance. An example of an initial representation for *break* is: ((+) (u -) (0)) break ((-) (u -) (0)), where (+) stands for active, (-) for passive, (u -) for indexing of the internal environment, (0) for duration. If the task is MT, the requirement for understanding could be shallow as pointed out by Wilks [7], although he did not discuss any dynamic grain adjustment. Contextual information can be conveyed by active features [5].

Following the boot-strapping principle, we are starting with 750 genus verbs in the defining word list of LDOCE, then gradually expanding them to all the verbs defined in LDOCE.

There are various subtasks associated with this work:

1. Dynamically adjusting classifications of relational concepts (mainly reflected by verbs): we are trying to get a set of core verbs as prototypes of classes based on certain statistics and genus verb sense nets (the latter is being constructed by G. Stein). A primary set of core verbs have been chosen, functional verbs are carefully prevented. The criterion for dynamically adjusting verb classes is: $C_j(d) = \{y : \|y-x\| \leq d, x \in C_j\}$, where C_j are core classes and $\|\cdot\|$ is defined as: $\|y-x\| = \min_i \{ i \text{ is the numbers of links on } P, P \text{ is any path connecting } x \text{ and } y \}$. We can select a reasonable distance for $C_j(d)$ by detecting slopes with points in the distribution of members. Classification can also be done within connectionist models.

2. From the prototypes, naive theories may be formed, and then converted into representations in the default layer.

3. Dynamic creation of Cases. Initial representations in the context layer may be ad-

justed and new Cases be created according to a set of contextual conditions (mainly when mismatches happen).

4. A set of rules can be constructed to get the conventional Cases for typical situations.

Many Case Theories are focused on verbs. In our situation, all the four major categories (verb, noun, adjective and preposition) must be paid enough attention to, since there are many verbs defined by verb phrases in LDOCE. e.g., a definition entry of verb *take* in LDOCE contains *get possession of*. In order to select a right Case frame and verb class for each verb, we need something beyond what we have presented although it does not conflict with what we have proposed and it is very plausible that the procedure used here may be adapted to establish Case frames for nouns, adjectives and prepositions. This task may be benefited from [2].

References

- [1] B. Atkins et al, *Explicit and Implicit Information in Dictionaries*, CSL Report 5, Princeton University, 1986.
- [2] R. Bruce and L. Guthrie, *Genus Disambiguation: A Study in Weighted Preference*, MCCS-91-207, CRL/NMSU, 1991.
- [3] C. Fillmore, *The Case for Case*, in *Universals in Linguistic Theory*, E. Bach and R. Harms (eds.), Holt, Rinehart, and Winston, 1968.
- [4] R. Schank, *Conceptual Information Processing*, North-Holland Publishing Co., 1975.
- [5] F. Weng, *A Three-Layer a posteriori Case Theory*, in preparation, 1991.
- [6] W. Wilkins, *Syntax and Semantics*, Academic Press, Inc., California, 1988.
- [7] Y. Wilks, *An Artificial Intelligence Approach to Machine Translation*, in *Computer Models of Thought and Language*, R.Schank and K.Colby (eds.), W.H.Freeman Co., 1973.