Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation

Nancy Ide and Jean Véronis

Laboratoire Parole et Langage
CNRS & Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 (France)
e-mail: {ide,veronis}@fraix11.univ-aix.fr

Department of Computer Science
Vassar College
Poughkeepsie, New York 12601 (U.S.A.)
e-mail: {ide,veronis}@cs.vassar.edu

Abstract. Machine-readable versions of everyday dictionaries have been seen as a likely source of information for use in natural language processing because they contain an enormous amount of lexical and semantic knowledge. However, after 15 years of research, the results appear to be disappointing. No comprehensive evaluation of machine-readable dictionaries (MRDs) as a knowledge source has been made to date, although this is necessary to determine what, if anything, can be gained from MRD research. To this end, this paper will first consider the postulates upon which MRD research has been based over the past fifteen years, discuss the validity of these postulates, and evaluate the results of this work. We will then propose possible future directions and applications that may exploit these years of effort, in the light of current directions in not only NLP research, but also fields such as lexicography and electronic publishing.

1 Introduction

The need for robust lexical and semantic information to assist in realistic natural language processing (NLP) applications is well known. Machine-readable versions of everyday dictionaries have been seen as a likely source of information for use in NLP because they contain an enormous amount of lexical and semantic knowledge collected together over years of effort by lexicographers. Considerable research has been devoted to devising methods to extract this information from dictionaries (see, for instance, [1,2,3,4,5,6,7,8,9,10,11]) based on the supposition that it is sufficient to form the kernel of a knowledge base that can be extended by utilizing information from other sources.

Interest in machine-readable dictionaries (MRDs) as a ready-made source of knowledge has waned somewhat in recent years. The number of papers on the topic at computational linguistics conferences and workshops (including the NewOED conference, which is devoted to the topic) is reduced, indicating either that extraction methods are well-established and robust (which is unlikely) or that research has turned to other areas. At the same time, the NLP community has turned its attention to corpora as a source of linguistic knowledge, evident both in the upsurge in the number of papers, journals, workshops, etc. dealing with corpora as a linguistic resource (e.g., the recent issue of *Computational Linguistics*, vol. 19, 1-2, 1993, devoted to corpus-based work, the workshop at ACL'93 on corpora, etc.) and in recent large-scale funding patterns (e.g., the European LRE program for corpora, ARPA's Linguistic Data Consortium in the U.S., etc.). It is clear that MRDs failed to live up to early expectations that they would provide a source of ready-made, comprehensive lexical knowledge. But does this mean that these many years of work on MRDs constitutes wasted effort? Does it mean that MRDs are conclusively unsuitable as a source for automatically building knowledge bases? In fact, while work on MRDs is eclipsed by the recent interest in corpora, no comprehensive evaluation of the value of MRDs as a knowledge source has been made in the light of the past ten or fifteen years' experience. Given the recent trend away from MRDs, as well as a greater understanding of what is needed in knowledge bases for NLP and how partial knowledge can be exploited, such an evaluation is timely in order to determine what may--or may not--be valuable to retrieve from MRD research.

This paper will first consider the postulates upon which MRD research has been based over the past fifteen years, discuss the validity of these postulates, and evaluate the results of this work. We will then propose possible future directions and applications that may exploit these years of effort, in the light of current directions in not only NLP research, but also fields such as lexicography and electronic publishing. We recognize a convergence of interests and goals between NLP and these other communities (cf. [12,13]), which may result in a benefit to all of them.

2 Thesis

MRD research has been based on two implicit postulates:

Postulate P1. MRDs contain information that is useful for NLP.

Postulate P2. This information is relatively easy to extract from MRDs.

The basis for postulate 1 is obvious when one considers a definition like the one for *fork* in Fig. 1, which identifies several semantic relations between *fork* and other lexical items that might be found in a semantic network. Access to this kind of information is often essential for many NLP tasks (sense disambiguation, PP attachment, etc.).

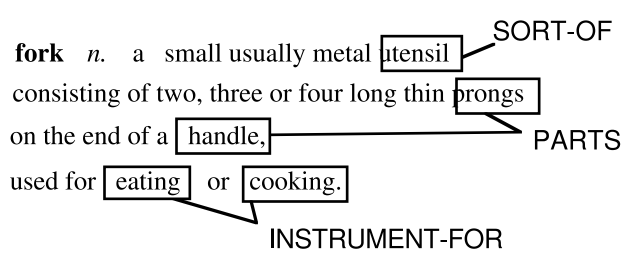

Fig. 1. Semantic information in definitions

Postulate 2, that most of this information is relatively easy to extract, was reinforced by work such as that described in [2] and [3], which proposed simple heuristics for automatically extracting hypernyms. These heurisitics exploit the fact that definitions for nouns typically give a hypernym term as the head of the defining noun phrase, as demonstrated in the following examples from the *Collins English Dictionary (CED)*:

dipper • a *ladle* used for dipping...
ladle • a long-handled *spoon*...
spoon • a metal, wooden, or plastic *utensil*...

The apparent success of extraction strategies such as this led to a flurry of activity in the area and the application of these techniques to various MRDs. Early results in MRD research were promising and led many to feel that MRDs would provide perhaps the most important source of information for building knowledge bases automatically (see, for instance, the position papers by Amsler and Boguraev at TINLAP-3 in 1987 [14, 15]).

3 Antithesis

However, despite early success in using information from machine readable dictionaries, it is still not clear that the results of early studies will scale up to enable automatically building full-scale knowledge bases from MRDs. In fact, the previous ten or fifteen years of work in the field has produced little more than a handful of limited and imperfect taxonomies, which hardly lives up to early expectations. In addition, there have been very few studies assessing the quality and usefulness of information extracted from MRDs, or showing how to systematically extract more complex information from definition texts.

In order to fully assess the value of MRD research to date, it is necessary to reconsider the fundamental postulates underlying this work by asking the following:

(1) How useful is the information in MRDs? In particular, is the information complete, coherent, and comprehensive enough to provide a basis for building knowledge bases? If not, what kind and how much of the information is missing?

(2) Is the extraction of information from MRDs as simple as applying strategies such as that of Chodorow, Byrd, and Heidorn?

3.1 Discussion of postulate P1

Although many studies boasted a high success rate in extracting information from MRDs, it has never been clear that extracted information is coherent or comprehensive enough to form a basis for knowledge bases. This is particularly true for information extracted from definition texts. Ide and Véronis [7] show that even in the most straightforward case (detection of hypernyms for concrete objects, kitchen utensils) 50-70% of the information is garbled in some way in five major English dictionaries. Similarly, it is not clear that the information in dictionaries is precisely what is needed to build knowledge bases. For example, Kilgariff [16] suggests that sense distinctions in the *LDOCE* (one of the most-used dictionaries in MRD research, which, as a learner's dictionary, is also one of the simplest) do not reflect actual use, and therefore may not form a viable basis for building knowledge bases for NLP. In addition, some types of knowledge simply do not exist in dictionaries.

Dictionary information is flawed. It is well known that information in the definition texts of dictionaries is often seriously inconsistent. Since a dictionary is typically the product of several lexicographers' efforts and is constructed, revised, and updated over many years, inconsistencies in the criteria for constructing definition texts necessarily evolve. In addition, space and readability restrictions as well as syntactic restrictions on phrasing may dictate that certain information is unspecified or left to be implied by other parts of the definition.

A pervasive problem in automatically extracted hierarchies is the attachment of terms too high in the hierarchy, which occurs in 21-34% of the definitions in the sample from the five dictionaries cited in [7]. For example, while *pan* and *bottle* are *vessels* in the *CED*, *cup* and *bowl* are simply *containers,* the hypernym of *vessel.* The problem of attachment too high in the hierarchy is compounded by the fact that it occurs relatively randomly within a given dictionary, which is evident in the hierarchies shown in Fig. 2. Because of inconsistencies such as these, semantic networks extracted from different dictionaries look very different, as demonstrated in the same figure.

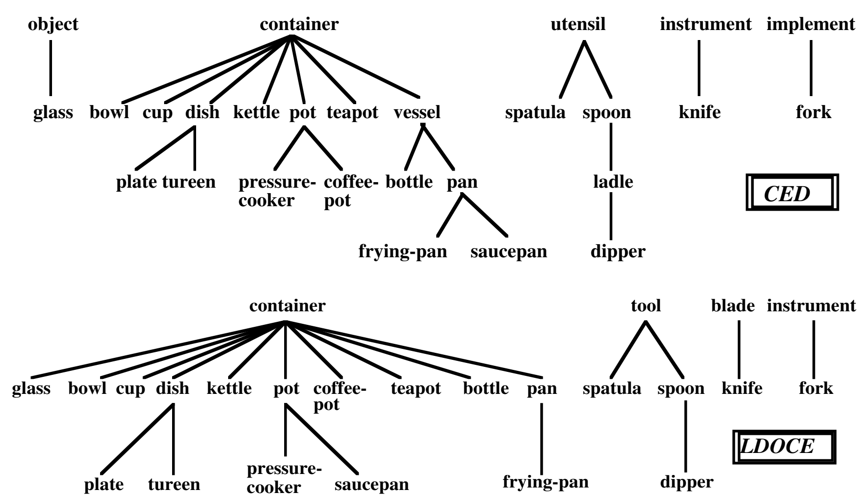

Fig. 2. Hierarchies from different dictionaries

In some cases, information is missing altogether. For example, the definition of *corkscrew* from *Webster's 9th*, "a pointed spiral *piece* of metal...", gives "piece" as the hypernym, which is clearly incorrrect. *Device,* which is the hypernym given in several other dictionaries, is better and should have been given if the dictionary were consistent.

Another pervasive problem with hypernyms generated from MRDs concerns information at the higher levels of the hierarchy, where terms tend to become more general and less clearly defined. For example, most people will agree on whether some object falls into the category *fork* or *spoon,* but there is much less agreement on what objects are *implements* or *utensils.* In addition, at the higher levels some concepts simply lack a term to designate them exactly. This lack of clear-cut terms for higher level concepts generates (at least) two phenomena in dictionary definitions. First, when higher level concepts are being defined, definitions are often circular. For instance, consider these definitions from the *CED*:

> **tool** • an *implement,* such as a hammer...
> **implement** • a piece of *equipment*; *tool* or *utensil.*
> **utensil** • an *implement, tool* or *container...*

Circular definitions yield hierarchies containing loops, which are not usable in knowledge bases. Second, in such cases definitions often give a list of head nouns separated by the conjunction "or", as in the definitions of *implement* and *utensil* above. In most cases none of the three alternatives is a true hypernym of the word being defined. Regarding them as such leads to other problems in the resulting hierarchy (Fig. 3): in the hierarchy produced from the definition of *utensil,* enumerating the paths upwards from *spatula* (defined as a *utensil*) leads to the conclusion that *spatula* is a kind of *container*, which is obviously incorrect. As demonstrated in [7] the introduction of "covert categories", that is, concepts which do not correspond to any particular word, can help to solve this problem, for instance, by introducting a covert category such as INSTRUMENTAL-OBJECT and using it as a hypernym for *tool, utensil, implement,* and *instrument*. However, the detection and creation of covert categories must be done by hand for the most part.

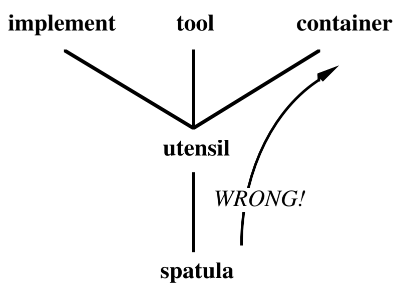

Fig. 3 : Problematic hierarchy

Despite irregularities of this kind, the specification of hypernyms is certainly more consistent in dictionaries than that of other semantic relations (e.g., parts, shape, color, smell, etc.), which is given in a much more random way. For example, in the definitions for *abricot* (apricot) and *pêche* (peach) from three major French dictionaries in Fig. 4, all of the dictionaries indicate that the pit of a peach is hard, but none specifies this property for the pit of the apricot (which is certainly just as hard). Information is given inconsistently even within a given dictionary, as shown when the definitions are presented in tabular form (Fig. 5).

	MICRO-ROBERT	HACHETTE	PETIT LAROUSSE
abricot	Fruit de l'abricotier, à noyau, à chair et peau jaune orangé.	Fruit de l'abricotier, d'une saveur délicate et parfumée, de couleur jaune rosé.	Fruit de l'abricotier, à noyau lisse, à peau et chair jaunes.
pêche	Fruit du pêcher, à noyau très dur et à chair fine.	Fruit comestible du pêcher, au noyau dur, à la chair jaune ou blanche, tendre et sucrée, à la peau rose et duveteuse.	Fruit comestible du pêcher, à chair juteuse et à noyau dur.

Fig. 4. Definitions for *abricot* (apricot) and *pêche* (peach)

	apricot			**peach**		
Property	**MR**	**HA**	**PL**	**MR**	**HA**	**PL**
sort-of fruit	yes	yes	yes	yes	yes	yes
parent-tree	yes	yes	yes	yes	yes	yes
has a pit	yes	no	yes	yes	yes	yes
form	no	no	no	no	no	no
color	no	no	no	no	no	no
texture	no	no	yes	no	no	no
hardness	no	no	no	yes	yes	yes
has flesh	yes	no	yes	yes	yes	yes
color	yes	no	yes	no	yes	no
texture	no	no	no	yes	no	no
hardness	no	no	no	no	yes	no
taste	no	yes	no	no	yes	no
has skin	yes	no	yes	no	yes	no
color	yes	(yes)	no	no	yes	no
texture	no	no	no	no	yes	no
hardness	no	no	no	no	no	no
etc.						

Fig. 5. Incoherence of properties in definitions

Information may be inappropriate. Some of the information in dictionaries is clearly not what is needed for NLP. For example, Kilgarriff [16] shows that in a sample of 83 words from the LOB corpus, 69 had at least two senses in LDOCE and were therefore ambiguous. However, Kilgarriff found that for 60 of these 69 words (about 87%), there was at least one usage in the LOB corpus which could not with any confidence be classified into a single sense. This occurred because, for example, more than one sense was near the meaning of the word as used in the corpus, or because no sense given in the dictionary applied.

MRD research has assumed for the most part that sense distinctions in dictionaries correspond to sense distinctions that apply in actual use, and therefore could provide the conceptual divisions that should appear in a knowledge base.

However, apart from distinctions between homographs, it is not clear that this assumption holds. The differences in the level of detail and, occasionally, in the ways lines are drawn between senses when one moves from one dictionary to another, already show that sense distinctions in dictionaries are not definitive; studies such as Kilgarriff's bring this fact into focus in relation to real language use, which is obviously the ultimate concern of NLP.

Missing types of information. Some types of information that must be included in a knowledge base are clearly not included in MRDs, in particular, broad contextual or world knowledge. For instance, it is interesting to note that there is no direct connection drawn between *lawn* and *house,* or between *ash* and *tobacco* in the *Collins English Dictionary,* although it is clear that this connection is part of human experience. The sense disambiguation strategy described in [17], which applies a connectionnist approach and relies on such connections between words in definition texts, fails in these cases as a result.

3.2 Discussion of postulate P2

The work of [3] and others made the extraction of semantic information from MRDs appear simple. However, claims of high success rates (often, 98% etc.) were misleading, since "success" (in Chodorow *et al.*'s case [7]) meant finding the *head* of a definition. This did not mean that this head was in fact an appropriate hypernym. In addition, by far the greatest success was achieved for simple semantic information, notably hypernyms. The extraction of other kinds of semantic information proved to be much more problematic due to far greater inconsistencies in the ways it was specified, often demanding relatively sophisticated parsing of the definition text . The following outlines the various sources of difficulty in extracting information from MRDs (see [8, 18, 19, 20]).

Physical formats. MRDs typically come to researchers in unusable formats--notably, in the form of typesetter tapes from publishers. To make the MRD usable for research, considerable effort was often required, and in fact the translation of MRDs in typesetter format to something more usable has become an area of study in itself. As a result of ambiguities and inconsistencies in typesetter formats, parsing typesetter tapes requires developing a complex grammar of entries (see, for example, [21]). Even with this, problems still arise because conventions are inconsistent. For example, in the *CED*, the entry *Canopic jar, urn* or *vase* must be interpreted as (*Canopic jar*) or (*Canopic urn*) or (*Canopic vase*), whereas the entry *Junggar Pendi, Dzungaria,* or *Zungaria,* which has the same structure, must be interpreted as (*Junggar Pendi*) or (*Dzungaria*) or (*Zungaria*). Because of the inconsistency, fully automated procedures cannot determine the appropriate interpretations.

Most of this work is far outside the realm of NLP research, and in general it is time-consuming and without great intellectual interest. Thus it diverts resources from more central NLP research. As a result, only a handful of dictionaries are available in a usable format, mainly in English. Now that the magnitude of the task is obvious, researchers may be reluctant to start similar work on dictionaries in other languages.

Incoherence of metatext. The difficulty of extracting information arises from inconsistencies in the way information is specified, in particular, variations in definition *metatext* (that is, phrases in definition texts that express semantic relations, such as "used in *V*-ing" for instrument, "consisting of a *N*" for parts, etc.). For example, consider the following from the *CED,* in which the fact that a handle is a part-of a utensil is expressed in vastly different ways:

> **jug** • a vessel...usually <u>having</u> a handle...
> **kettle** • a metal container <u>with</u> a handle...
> **ladle** • a <u>long-handled</u> spoon...
> **corkscrew** • a device...<u>consisting of</u> a pointed metal spiral <u>attached to</u> a handle...
> **fork** • a small usually metal implement <u>consisting of</u> two, three, or four long thin prongs <u>on the end of</u> a handle...
> **knife** • a cutting instrument <u>consisting of</u> a sharp-edged often pointed blade of metal <u>fitted into</u> a handle...
> **basket** • a container...often <u>carried by means of</u> a handle or handles.

This small sample demonstrates that there are often a handful of metatextual phrases signalling particular relations ("having a", "with a", "consisting of", etc.) that can be detected with simple Chodorow-like pattern matching techniques. Even this simple case is problematic; consider

> **lug** • a box or basket for vegetables or fruit <u>with</u> a capacity of 28 to 40 pounds.

where the noun (*N*) in the phrase "with a *N*" is not a part, as it is in the earlier examples.

However, things are rarely even this straightforward. The definitions of *corkscrew, fork, knife,* and *basket* show that there is virtually an open-ended set of metatextual phrases to specify that a handle is a part of some object, for which no pattern can be devised. Bearing in mind that this example shows only a single case (handle) over a handful of definitions in a single dictionary, it is clear that the amount of work required to determine the possible variants could easily exceed that required to construct the corresponding knowledge base by hand.

The bootstrapping problem. It is common in MRD research to use sophisticated syntactic parsers (for example, the Linguistic String Parser as in [22]) to analyze definition texts. But even for the simple examples given above, more than syntactic analysis is required. For example, to determine that a handle is a part-of a basket from the *CED* definition, not only syntactic parsing but also lexical information is needed to differentiate "carried by means of a handle" from "cooked by means of steam". Similarly, in the definition of *ladle,* lexical information would be required in order to determine that the adjective "long-handled" specifies a part of a ladle. In some cases, the resources of a full knowledge base are required to understand the specification. For example, to differentiate "carried by means of a handle" from "carried by means of a wagon", world knowledge is required to determine that "wagon" is not a part of the object carried, although "handle" most likely is.

In addition, in order to create knowledge bases from MRDs, it is necessary to apply sense disambiguation to the words in the dictionary itself as a prior step. Without prior sense disambiguation, automatically extracted hierarchies are necessarily *tangled* [1], because many words are polysemous. For example, in the *CED*, the word *pan* has the following senses (among others):

pan[1] **1.a** a wide metal *vessel*… [CED]
pan[2] **1** the *leaf* of the betel tree... [CED]

The *CED* gives *pan* as the hypernym for *saucepan,* which taken together yields the hierarchy in Fig. 6. The undisambiguated hierarchy is unusable because, following the path upwards from *saucepan*, we find that *saucepan* can be a kind of *leaf,* which is clearly erroneous. Sense disambiguation is a complex task requiring possibly substantial knowledge resources.

All of this means that in order to create resources for use in NLP from MRDs, it is necessary to have full NLP capabilities--including full knowledge bases--already at hand. This is clearly circular, and has led some researchers to attempt MRD analysis using only minimal pre-existing resources that can be constructed by hand, and bootstrapping as MRD analysis proceeds. However, it is becoming clear that the difficulties are so considerable that such methods are unlikely to succeed; indeed, to date, none has been convincingly demonstrated.

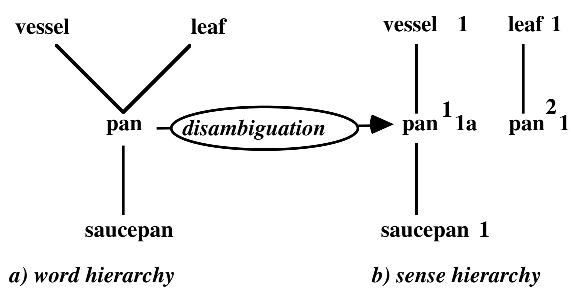

Fig. 6 : Sense-tangled hierarchy

4 Synthesis

The remarks in the preceding sections lead to the conclusion that the two postulates of MRD research were erroneous. The information in MRDs is probably not consistent or complete enough to provide a substantial basis for a knowledge base. In addition, given that the amount of work required to extract the information that does exist in MRDs has now been demonstrated to be extensive (even requiring the existence of the knowledge bases it is intended to create), the return on investment is clearly very low.

Therefore, does the past fifteen years of MRD research constitute wasted effort? If the mostly automatic extraction of near-perfect knowledge is the goal, the answer is certainly "yes". But this assumption reveals a broader postulate than the two cited in section 2 underlying MRD research:

Postulate P0. Large knowledge bases cannot be built by hand.

It is not at all sure that this postulate is valid, given, for example, the work of the CYC project [34], and the fact that lexicographers have been creating knowledge bases by hand for over 200 years. If this broader postulate is abandoned, we can more realistically assess the contribution of MRD research. It is clear that MRDs contain useful data--but most of it is probably usable only with possibly substantial by-hand massaging, requiring human judgement to be incorporated into useful knowledge bases. In addition, it is the trend in NLP research (possibly in part because of the experience with MRDs) to consider that no single source (dictionaries, corpora, etc.) could provide all or even most of the knowledge required for NLP. Therefore, it is now widely recognized that knowledge base construction requires combining information from multiple sources (section 4.1). Clearly, coupled with information from other sources and subjected to by-hand amelioration, information extracted from MRDs is a valuable resource for building knowledge bases.

However, there are other contributions of MRD research that may be less well-recognized. In fact, MRD research has contributed significantly to several areas, in particular, the development of encoding and database models for dictionaries and other textual data (section 4.2), the assessment of the kind of knowledge needed for NLP (section 4.3), and the consideration of pure associational information for various NLP tasks (section 4.4).

Probably the most important contribution of MRD research is the fostering of a convergence of interests among the fields of NLP, lexicography, and electronic publishing--a convergence which MRD research will certainly continue to feed and develop. This convergence promises to benefit NLP research as well as lexicography and electronic publishing (section 4.5).

4.1 Combination of knowledge sources

It is now becoming clear that merging information from multiple sources is essential to the process of creating knowledge bases. One such possibility involves the use of information from several dictionaries, since although information derived from individual dictionaries suffers from incompleteness, it is extremely unlikely that the same information is consistently missing from all dictionaries. It is therefore possible to use information from several dictionaries to fill in information which is missing or faulty in one or more others. For example, in a small experiment, merging multiple dictionaries produced highly encouraging results: in a merged hierarchy created from five English dictionaries, the percentage of problematic cases was reduced from 55-70% to around 5% [7]. By-hand work is still required, but merging can substantially improve the quality of the extracted information.

It is now widely recognized that combining information extracted from MRDs with information provided by corpus analysis is a fruitful means to fill out knowledge bases, since corpora can provide information such as common collocates, proper nouns, role preference information, frequency of use and similar statistics, etc. However, with corpora as with MRDs, fully automatic extraction is not likely.

We foresee that the creation of knowledge bases in the future will be accomplished by giving the human knowledge-base-creator access to multiple resources, including MRDs and corpora, together with tools to extract different kinds of information and combine it more or less by-hand. Indeed, projects to develop workstations for this kind of work are already underway. This information will be widely varied in kind, including both detailed linguistic information as well as statistics, associational links, etc.

4.2 Physical organization of dictionaries

MRD research has necessarily involved considerable work on rendering dictionaries into a usable format for extraction of information. To this end, an encoding format for MRDs has been developed under the aegis of the Text Encoding Initiative (see [23, 24, 25, 26]) which can be applied across mono- and bi-lingual western dictionaries. Such a format must necessarily be both general enough to be applicable across different dictionaries, whose structures often vary widely, and at the same time capture the fundamental structural principles (e.g., hierarchical structure, factoring of information) that underlie dictionaries. A common encoding format enables the application of common software and hence the reusability of MRDs, and is extremely useful in the publishing industry for rendering in-house data in common formats which are directly suitable for typesetting, generation of dictionaries in different forms (e.g., concise, learner's), etc.

The development of an encoding format suitable for MRDs demands identification of the dicionary entry's constituent elements as well as a deep understanding of the structural principles underlying dictionaries. Thus a related problem is the determination of a database model suitable for representing the information in MRDs. Lexical data, as is obvious in any dictionary entry, is much more complex than the kind of data (suppliers and parts, employees' records, etc.) that has provided the impetus for most database research. Therefore, classical data models (for example, relational models) do not apply very well to lexical data, although several attempts have been made (see for example [6, 27]). Ide, Véronis, and Le Maitre [28] have proposed an alternative feature-based model for lexical databases, which allows for a full representation of sense nesting and defines an inheritance mechanism that enables the elimination of redundant information. The model has been implemented in an object-oriented DBMS [29]. This and other similar work continues to feed the development of database models to represent lexical data and textual information, which is becoming an increasingly active area of research in database design.

4.3 Assessment of the needs of NLP

MRD research to date has provided the basis of an assessment of the kind of knowledge that is needed in NLP. In particular, in comprehensively examining the information presented by lexicographers who had no conscious intention to provide a fully systematic set of semantic specifications, numerous subtleties concerning semantic distinction and semantic relations have become more clear. The covert category problem and the attendant thinking about overlapping meanings, circularity, etc., described in section 3.1.1 provides one simple example.

The exact nature and kind of information required for various NLP tasks has not been fully explored. For example, even in the case of taxonomies, it is not always clear what must be included: for instance, we know that some inheritance mechanisms are needed for parsing (e.g., for successful PP attachment in "I ate a trout with bones", it is necessary that *trout* inherits the feature *has-bones* from *vertebrate*)--but are there cases where more precise or different kinds of information are necessary? Is the broad semantic information in a dictionary sufficient? Too much? It is difficult to draw a precise taxonomy in many cases (e.g., it is very difficult to determine whether a given item is a *pot*, a *pan,* or both), and yet humans easily understand sentences containing words for which the taxonomic relations are unclear. This suggests that the kind of precision NLP researchers have traditionally sought in knowledge bases may be unnecessary in some cases. We can ask if very different kinds and amounts of information required for different NLP tasks, and if so, it will be essential to precisely identify these differences. Some studies have attempted to assess in detail the kinds of knowledge required for given NLP tasks (see, for example, [30]); such studies provide a start, but it is clear that more consideration of the exact requirements of various NLP tasks needs to be done.

4.4 Exploitation of associational information

Work on extracting semantic information from definition texts involves an attempt to identify not only which words are related, but also the nature of the relationship. However, several researchers have used MRDs as a source of *associational* information indicating which words are related, without regard for the nature of the relationship, to perform tasks such as sense disambiguation, topic identification, etc. The fundamental assumption underlying these studies is that words in a definition are closely related semantically to the defined word. Such studies have identified senses by counting overlaps between words in the definition texts of different senses and surrounding words in context [31, 9] and generated lists of topic or key words by extracting the not only the most frequent words in a text, but also the most frequent words in their definitions [32]. Other studies have employed the same principle by building associational networks on the basis of the implicit association between headword and words in the definition text for sense disambiguation ([17, 33]).

The work which has utilised the word associations implicit in a dictionary's structure has shown considerable success, although like much NLP research these studies have typically involved small-sized experiments, and it is not clear that the methods will scale up to real-size data. However, at the very least, this work demonstrates that the associational information in MRDs--which is trivial to extract since no complex processing is required--is potentially valuable for use in NLP.

4.5 Synergy among NLP, lexicography, electronic publishing

One of the most promising possibilities for the future of MRD research results from a merging of interests among NLP researchers, lexicographers, and electronic publishers. Lexicographers, possibly as a result of MRD research, are increasingly interested in creating lexical data bases containing the kinds of information that NLP research had hoped to extract from MRDs; some lexicographers are explicitly concerned with creating NLP-like knowledge bases [12, 13]. Creation of such databases would in turn provide NLP with more of the resources it needs. In addition, electronic publishing has made possible the creation of commercially available, hypertext-like dictionaries which would include information and facilities well beyond that of print dictionaries. This sort of product is likely to explode on the market in the near future, and in fact a few such dictionaries (Sony Data Discman, Larousse, etc.) already exist.

So far, computerization has been applied to lexicography in only limited ways; the COBUILD project [35] was one of the first to utilize computers to exploit corpora in the creation of lexical entries, and most dictionary publishers now create dictionaries by first creating in-house databases--although such databases typically contain only gross distinctions among information fields (orthographic form, pronunciation, part-of-speech, etymology, definition text, etc.). Almost no work has been done to improve definition texts themselves or to systematize semantic information (apart from occasional attempts such as the *LDOCE* semantic codes).

Computerization of dictionary-making at the semantic level could involve the following:

(1) *the creation of explicit semantic links* (hypernym, part, color, etc.) between words or entries. This would be especially useful for creating electronic (hypertextual) dictionaries. Such links could lead to the development of precise templates for classes of objects, etc. (e.g., Fig. 7).

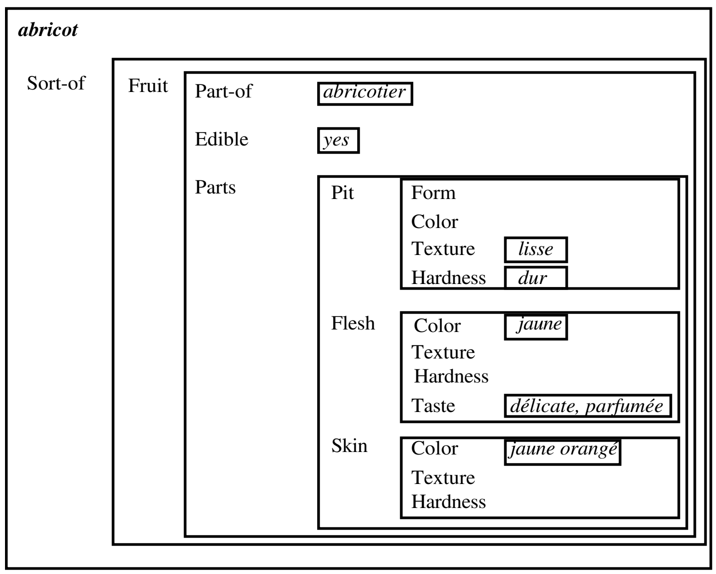

Fig. 7. Template for dictionary entries (fruits)

So far, navigation and query in electronic dictionaries is rather limited and relies on user's judgement and understanding of definitions to be usable. However, an explicit semantic net underlying the dictionary could be very useful for nativation and query. For example, we can envision display to varying levels of detail, depending on user preference, of the information in the template for *fruit*, and even user navigation within the template (click on PARTS-OF "apricot" and get "pit," "flesh," "skin," etc.; click on "pit" and get the properties of apricot pit, etc.). Information could be linked to images and sounds, and displayed in template form; or definitions and sub-definitions could be generated in natural language, in any form (concise, learner's version, full, etc.).

(2) *ensuring consistency of the content of entries.* The templates created for different classes of concepts could be used to ensure that the information given for each entry when it is appropriate to do so. This could eliminate the kinds of inconsistencies demonstrated in the entries for *apricot* and *peach* given in section 3.1.1.

(3) *ensuring consistency of metatext.* As outlined above in section 3.2.2, MRD research has revealed that metatext in dictionaries is highly variable and relatively inconsistent, and, by compiling lists of metatextual specifications for various relations, has identified both the sources and potential solutions to the problem of

inconsistency. We can even imagine that metatext could be automatically generated from templates of the type described above.

(4) *ensuring consistency of sense division.* Lexicographers have sought means to remove the arbitrariness of sense divisions. Computerization has already been helpful, for example, by automatizing and systematizing the use of corpora as a source of information about word senses. Knowledge bases could take this systematization even farther. For instance, a dictionary might define *cup* as "**1.** a container for liquid... **2.** its content", *bowl* as "**1.** a container...; its content", but *glass* only as "**1.** a container...", which ignores the metonymic use of glass ("its content") and is therefore inconsistent. An electronic database containing explicit marking of metonymic links could enable checking that metonymic use is specified where necessary, and in a consistent format.

5 Conclusion

The false expectation that large knowledge bases could be generated automatically from MRDs has led to a perception that the past fifteen years of MRD research has failed to meet the original goals. Indeed, it seems to be the case that large-scale knowledge bases will be built using information from multiple sources, and will require human involvement. From this perspective, it is clear that while they are not the exclusive resource they may have been originally thought to be, MRDs have something to contribute to the creation of knowledge bases. In some instances that contribution is not what had originally been expected, as evident in the fact that MRDs have been found to contain a vast bank of associational information that is useful in many NLP tasks.

It is also clear that MRD research has in fact contributed to other NLP goals. In particular, it has contributed to our understanding of the nature, kinds, and role of semantic information in the processing of natural languages, as well as to the increasingly obvious idea that widely varying types of information--several in addition to the traditionally accepted set--are needed for NLP. This may in turn lead to more systematic assessment of the needs of various NLP tasks, an area which deserves serious attention.

The most promising avenue of activity, however, involves collaboration between the NLP community and lexicographers and electronic publishers. The two communities are already beginning to work with one another; one clear example is a recent survey sent to NLP researchers from Longman publishers, asking for their input in devising new database versions of the *LDOCE.* Obviously, collaboration--both in terms of shared information and shared effort--can benefit both communities. The result could be better dictionaries and exciting new possibilities for electronic, hypertextual dictionary databases, as well as a wealth of material useful for NLP.

References

1. R. A. Amsler: The structure of the Merriam-Webster Pocket Dictionary. Ph. D. Dissertation, University of Texas at Austin, (1980)
2. N. Calzolari: Detecting patterns in a lexical data base. Proceedings of the 10th International Conference on Computational Linguistics, COLING'84, 170-173 (1984).

3. M. S. Chodorow, R.J. Byrd, G. E. Heidorn: Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics, Chicago, 299-304 (1985).
4. J. Markowitz, T. Ahlswede, M. Evens: Semantically significant patterns in dictionary definitions. Proceedings of the 24th Annual Conference of the Association for Computational Linguistics, New York, 112-119 (1986).
5. R.J. Byrd, N. Calzolari, M.S. Chodorow, J.L. Klavans, M.S. Neff, O. Rizk: Tools and methods for computational linguistics. Computational Linguistics, 13, 3/4, 219-240 (1987).
6. J. Nakamura, M. Nagao: Extraction of semantic information from an ordinary English dictionary and its evaluation. Proceedings of the 13th International Conference on Computational Linguistics, COLING'88, 459-464 (1988).
7. N. Ide, J.Véronis: Refining taxonomies extracted from machine-readable dictionaries. In: S. Hockey, N. Ide (eds.): Research in Humanities Computing 2. Oxford University Press (in press).
8. J. Klavans, M. Chodorow, N. Wacholder: From dictionary to knowledge base via taxonomy. Proceedings of the 6th Conference of the UW Centre for the New OED, Waterloo, 110-132 (1990).
9. Y. Wilks, D. Fass, C. Guo, J. MacDonald, T. Plate, B. Slator: Providing machine tractable dictionary tools. Machine Translation 5, 99-154 (1990).
10. F. Pigamo: Outils de traitement sémantique du langage naturel. Thèse de l'Ecole Nationale Supérieure des Télécommunications, Paris, 242pp. (1990).
11. A. Alonge: Analysing dictionary definitions of motion verbs. Proceedings of the 15th International Conference on Computational Linguistics, COLING'92 1315-1319 (1992),.
12. R. Martin: Inférences et définition lexicographique. Colloque "Lexique et Inférences", Metz, (1991).
13. P. Procter: Cambridge Language Survey: The development of a non-language specific semantic coding system using multiple inheritance. Paper presented at International Workshop of the European Association of Machine Translation, "Machine Translation and the Lexicon", Heidelberg, 26-28 (April 1993).
14. R.A. Amsler: Words and worlds. Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing (TINLAP-3). Las Cruces, NM (1987).
15. B.K. Boguraev: The definitional power of words. Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing (TINLAP-3). Las Cruces, NM, 11-15 (1987).
16. A. Kilgarriff: Dictionary word sense distinctions: An enquiry into their nature. Computers and the Humanities, 26 (5-6), 365-388 (1993).
17. J. Véronis, N.M. Ide: Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. Proceedings of the 14th International Conference on Computational Linguistics, COLING'90, Helsinki, 2, 389-394 (1990).
18. K. Jensen, J.-L. Binot: Disambiguating prepositional phrase attachements by using on-line dictionary definitions. Computational Linguistics 13, 3-4, 251-260 (1987).
19. S. Montemagni, L. Vanderwende: Structural patterns vs. string patterns for extracting semantic information from dictionaries. Proceedings of the 15th International Conference on Computational Linguistics, COLING'92, 546-552 (1992).

20. Y. Ravin: Disambiguating and interpreting verb definitions. Proceedings of the 28th Annual Conference of the Association for Computational Linguistics, Pittsburgh, 260-267 (1990).
21. M.S. Neff, B.K. Boguraev: Dictionaries, dictionary grammars and dictionary entry parsing. Proceedings of the 27rd Annual Conference of the Association for Computational Linguistics, Vancouver, 91-101 (1989).
22. T. Ahlswede, M. Evens, K. Rossi: Building a lexical database by parsing Webster's Seventh Collegiate Dictionary. Proceedings of the 2nd Annual Conference of the UW Centre for the NewOED, Waterloo, Canada, 65-76 (1985).
23. R.A. Amsler, F.W. Tompa: An SGML-based standard for English monolingual dictionaries. Proceedings of the 4th Annual Conference of the UW Centre for the New Oxford English Dictionary, Waterloo, Ontario, 61-80 (1988).
24. N. Ide, J. Véronis: Print dictionaries, TEI Working Paper AI5 D17, Distributed by the Text Encoding Initiative. Compter Center, University of Illinois at Chicago, 60pp. (1992).
25. N. Ide, J. Veronis, S. Warwick-Armstrong, N. Calzolari: Principles for encoding machine readable dictionaries, EURALEX'92 Proceedings, H. Tommola, K. Varantola, T. Salmi-Tolonen, Y. Schopp, eds., in Studia Translatologica, Ser. a, 2, Tampere, Finland, 239-246 (1992).
26. N. Ide, J. Véronis: Encoding dictionaries. Computers and the Humanities 29, 1-3 (to appear).
27. M.S. Neff, R.J. Byrd, O.A. Rizk: Creating and querying lexical databases. Proceedings of the Association for Computational Linguistics Second Applied Conference on Natural Language Processing. Austin, Texas 84-92 (1988),.
28. N. Ide, J. Le Maitre, J. Veronis: Outline of a model for lexical databases. Information Processing and Managment 29, 2, 159-186 (1993).
29. J. Le Maitre, N. Ide, J. Véronis: Deux modèles pour la représentation des données lexicales et leur implémentation orientée-objet. Actes des 9èmes Journées Bases de Données Avancées, Toulouse, 312-331 (1993).
30. S.W. McRoy: Using multiple knowledge sources for word sense discrimination. Computational Linguistics 18, 1, 1-30 (1992).
31. M. Lesk: Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. Proceedings of the 1986 SIGDOC Conference (1986).
32. N. Ide, J. Véronis: Caught in the web of words: Using networks generated from dictionaries for content analysis. Paper presented at ACH/ALLC'91 Joint International Conference, Tempe, Arizona (1991).
33. N. Ide, J. Véronis: Very large neural networks for word-sense disambiguation. 9th European Conference on Arfificial Intelligence, ECAI'90, Stockholm, 366-368 (1990).
34. D.B. Lenat, M. Prakash, M. Shepherd: CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. AI Magazine 7 (4), 65-85 (1986).
35. J.M. Sinclair: An account of the COBUILD project. London: Collins ELT 1987.