Caroline Barrière & Fred Popowich, Simon Fraser University, Canada

Building a Noun Taxonomy from a Children's Dictionary[1]

Abstract

This paper explores noun taxonomy extraction from a children's first dictionary. We suggest the use of such a dictionary because of its particular structure oriented toward the learning by children of a basic vocabulary used in everyday life. We discuss different ways of analyzing the dictionary definitions to find taxonomic links and we describe the implementation of some knowledge extraction and taxonomy construction techniques.

1. Introduction

Machine Readable Dictionaries (MRDs) contain much lexical information that would be useful to a Natural Language Processing (NLP) system, and a great deal of work has been done on knowledge extraction from such dictionaries (Amsler 1980, Calzolari 1984, Chodorow and al. 1985, Byrd and al. 1987, Wilks and al. 1993). The most popular structure extracted from dictionaries is the type hierarchy or taxonomy of nouns. Work on taxonomy extraction has been mostly done on dictionaries like the Webster 7 (an adult dictionary), or the LDOCE (a learner's dictionary), that are available in electronic versions. Here, we investigate noun taxonomy extraction from a children's first dictionary, the American Heritage First Dictionary (AHFD). During the process, we compare the definitions and the resulting classification of nouns between the AHFD and an adult's dictionary.

2. Using a children's first dictionary

This paper is part of a larger project that aims at building a Lexical Knowledge Base (LKB) containing information about how words are used, how they relate to each other and how they are put in context. Any dictionary contains a lot of lexical information that could be put in the LKB. But where is the best place to start? In this research, we will argue that a children's first dictionary is in fact a good starting point.

The AHFD, which contains 1800 entries, is addressed to children of age 6 to 8. It is made for young people who are learning the structure and the basic vocabulary of their language. The AHFD tries to give the most common definition of a word. In comparison, an adult's dictionary is more of a reference tool, which assumes knowledge of a large basic vocabulary. The adult's dictionary can be really specific, giving multiple senses and usages of words in different situations. For example, here are the definitions of the word *handkerchief* from the AHFD and from the American Heritage Dictionary (AHD):

[**AHFD:**]	A handkerchief is a piece of cloth. You put it over your nose when you sneeze. Many handkerchiefs are white.
[**AHD:**]	A small square of cloth used in wiping the nose, mouth, etc.

When defining a word, the AHFD always puts the defined word into a complete sentence. The sentences used are usually short and express one specific idea. The first or two sentences define the word, and then one or two sentences are used as typical situations involving the young reader.

The dictionary also has the property of defining every noun using other nouns and verbs that are themselves defined. Only a few exceptions use a noun usage of a defined verb. For example, *joy* is defined in terms of a *feeling*, and *feeling* itself is never defined, but the verb *feel* is. This property of having all words expressed in terms of other defined words will allow us to create a closed LKB as a core which could later grow from information extracted from other dictionaries or text corpora.

3. Extracting a noun taxonomy from the AHFD

Since most research on taxonomies has been done on nouns, we produced an electronic version of the AHFD[2] containing all 1035 noun definitions, leading to a total of 1117 wordsenses. Most definitions are written as a **genus** and a **differentia**. The genus specifies the class in which to put the noun, and the differentia specifies how different that noun is from the other nouns in the class. Here are two examples:

1. A **hospital** is a large building.
2. An **apple** is a kind of fruit.

As we look at different definitions, we find patterns that are used over and over and that correspond to different semantic relations. Those patterns are called defining formulas (Markowitz and al. 1986, Ahlswede and Evens 1988). Here, we have two defining formulas, <N *is a* {*adj**} N> (example 1) and <N *is a kind of* {*adj*} N> (example 2) leading to a IS-A relation. For each example, we can establish the IS-A relations between *hospital/building, apple/fruit*. In the AHFD, there are 586 instances of the pattern *is a/an*, and 151 instances of the pattern *is a kind of*.

As we build the hierarchies of IS-A relations, going upward, at the highest level we will find a circularity. The whole taxonomy will be a forest with multiple trees, each of which having at its root a group of words defined through a loop. This loop contains a group of synonyms. For example, part of our taxonomy extracted from the AHFD would be a tree having at its root the word *person*. *Person* is defined in terms of being a *man, woman, boy or girl* and all those lexical units have *person* as the genus of their definition.

In most research on information extraction from MRD, the taxonomy is extracted via the genus/differentia definition structure. But there should be other ways to find taxonomic links. We explore two other ways here, where it is possible to find some classes through a more detailed examination of the definitions.

3.1 Covert categories

Cruse (1986) introduced a notion of unlabeled categories that can be found using a sentence frame containing a variable X where we determine a list of items that X could be. For example, given the sentence frame *John looked at the X to see what time it was*, we could generate the list *clock, watch, alarm clock* as possible values for X. Cruse calls these categories with no names, but for whose existence there is definite evidence, **covert categories**.

Covert categories are often present in the AHFD, either because the label is not part of the vocabulary we want to teach the child, or because the label doesn't exist in the English language. Consider the *vehicle* category, which does have a label in English, but not (yet) in the children's world. In the AHFD, the concept of a vehicle is expressed through the action of *carrying*. The sentence frame could be *X carries/ carry people/loads*. Here are two of the 12 definitions from the AHFD that would match this sentence frame.

3. An **airplane** is a machine with wings that flies in the air.
   Airplanes carry people from one place to another.

4. **A boat** carries people and things on the water.

Overall, we find a category X including *airplane, balloon, boat, bus, camel, donkey, helicopter, ship, subway, train, truck, wagon.* All those lexical units have in common of *carrying people or loads.* Some of the items found were already assigned to the class *machine.* Some others like *camel, donkey* have *animal* as their class. But others didn't have any category assigned like *boat, train, wagon.*

3.2 Sets

Besides using sentence patterns, the discovery of sets of related words can also suggest the existence of a covert (or non-covert) category. Consider a structure of the type *X and other Y* (Cruse 1986) which we also find in the AHFD.

5. **A closet** is a very small room.
   People keep clothes, shoes, and other things in **closets**.
6. **Juice** is the liquid inside foods.
   People drink the **juice** of apples, oranges, grapes,
   tomatoes, and many other fruits.

When the word after *other* is quite precise, it could be assigned as the appropriate superclass for the set of items. But often, that word is very general, like *other things.* Then, the grouping of the words still suggest the presence of a superclass, but it could be either a covert category, or a category for which a name has to be found through further analysis of the dictionary.

4. Implementation

In this section we show how using Conceptual Graphs (CGs) (Sowa 1984) as our knowledge representation formalism, enables us to extract the classes for words that have their definition in the conventional genus/ differentia form, as well to find the covert categories expressed through phrasal patterns and the subclasses found as part of sets.

CGs are a logic-based representation formalism with a close mapping to natural language. They include concepts and the relations between them. Here is a sentence with its corresponding CG.

```
John eats the soup     [eat]->(agent)->[person:John]
with a spoon                ->(object)->[soup]
                            ->(instrument)->[spoon]
```

Our implementation is based on a CG environment called CoGITo (Haemmerle 1995) which allows us to define our graphs, maintain the type hierarchy, and perform some graph matching operations needed to compare the multiple graphs. Here we will focus on how the CGs associated with a definition can be used to modify the noun taxonomy. The process of transforming AHFD definitions into CGs is discussed in (Barrière and Popowich 1996).

The CG generated from a parsed definition may be transformed by a set of Semantic Relation Transformation Rules (SRTRs). Those rules are developed by finding patterns, or *defining formulas* in the AHFD which lead to specific semantic relations (Ahlswede and Evens 1988, Dolan and al. 1993). Table 1 shows some examples of SRTRs. Applying the SRTRs at the graph level instead of doing pattern matching at the sentence level makes it much easier to account for all possible pattern variations (Montemagni and Vanderwende 1992).

Pattern	Graph before SRTR	New Graph
A is a B	[is]->(object)->[B] ->(agent)->[A]	[A]->(is-a)->[B]
A is a kind of B	[is]->(object)->[kind]->(of) >[B] ->(subject)->[A]	[A]->(is-a)->[B]

Table 1. Semantic Relation Transformation Rules

To actually construct the noun hierarchy, we start with an initial hierarchy having only one level; all nouns are subclasses of "something". As we look at each definition, and generate graphs of the type [A]->(is-a)->[B], we can modify the hierarchy and place A as a subclass of B. As we do so for the whole dictionary, the hierarchy builds up.

4.1 Covert categories

For building the taxonomy automatically, finding the covert categories requires more work than finding the genus of a definition. The categories are hidden through phrasal patterns which are repeated among many definitions with maybe some variants. Those patterns are not known in advance and have to be found through comparisons of sentences.

As those patterns are usually centered around a verb, we decided to take each verb (except verbs *be*, *have*) and build a list of possible immediate relations for each of them. To do so we build a general graph (G1) around a verb. For example with the verb *carry*, we have:

G1 = [A]<-(r1)<-[carry]->(r2)->[B]

Using graph matching techniques, we project G1 onto all the graphs constructed from the dictionary definitions. As a result, we will have a list of all projections; meaning all subgraphs that are more specific than G1. We find nine occurrences of the projection where r1 is specialized to agent, r2 to object and B to people:

[A]<-(agent)<-[carry]->(object)->[people]

For our initial experiments, when the number of occurrences of a projection exceeds five, we use it to define a covert category. We make use of the λ-abstraction mechanism of CGs to label and define a new type concept. For our covert category, we generate a type concept named label-1 (it could later correspond to an existing word, since we saw that some words are not yet part of the child's vocabulary) and assign it a λ-abstraction given by the projection we found.

label-1(*Y) is [*Y]<-(agent)<-[carry]->(object)->[people]

We can then update the hierarchy and put all the nouns that could replace the coreference *Y under the type label-1.

4.2 Sets

Sets are easier to deal with as they correspond to a simple graph matching. We need to find a pattern like

```
[A]->(and)->[B1]
  ...
  ->(and)->[Bn]
  ->(and)->[C]<-(modif)<-[other]
```

The relation *modif* corresponds to any adjective. In our case, all type concepts A, B1...Bn become subclasses of C.

5. The outsiders

In the AHFD, the classification of a noun into a group is not always its most important feature. In some definitions, the first sentence describing the genus is completely missing, or the genus assigned is so general that it is uninformative. The sentence explains the usage or purpose of the word, trying to find a characteristic that would be more essential to its understanding than trying to give a genus that might be confusing for a child.

A definition in which the genus/differentia pattern isn't present will put the noun in relation to other parts of speech, often as a case relation to a verb as its typical object, agent or instrument. In the AHFD, more than 10 definitions are of that type. The genus is replaced by an overgeneral term, like *something* or *what* and the focus is on the action. Therefore the noun defined becomes a case relation to the verb. In some cases, that noun can be a general concept and become the root of a tree (e.g. *food*). The global taxonomy becomes a forest including all those trees.

The adult dictionary, on the other hand, will usually try to find a genus to the expense of getting into complicated sentence structures, as well as sometimes finding obscure nominalizations. Here are some comparisons:

7a. **Food** is what people or animals eat. (AHFD)
7b. **Food(1):** A substance taken in and assimilated by an organism to maintain life and growth; nourishment. (AHD)
8a. **Sound** is anything that you hear. (AHFD)
8b. **Sound(1a):**A vibratory disturbance, with frequency in the approximate range between 20 and 20,000 cycles per second, capable of being heard. (AHD)

For our implementation, we are not updating the noun hierarchy but we can still assign a λ-abstraction to those nouns. With *food* and *sound*, we have the λ-abstractions:

food(*X) is [*X]<-(object)<-[eat]->(agent)->[person]->(and)->[animal]
sound(*X) is [*X]<-(object)<-[hear]->(agent)->[person]

6. Conclusion

In investigating the construction of taxonomic relations from dictionary definitions, we used the American Heritage First Dictionary (AHFD) for its simplicity, its emphasis on daily usage by giving examples, its restricted number of senses, giving the most common senses of a word. When we look at the simple definitions of the AHFD, it is amazing to see how much information they actually give through the usage and examples, and how this information is often what we mostly need to understand a non technical daily conversation.

Although the AHFD most often gives the taxonomic link using the genus/differentia structure, we have explored the use of covert categories to incorporate unlabeled classes into the taxonomy. Updating the noun hierarchy is done by graph matching operations on the Conceptual Graphs corresponding to the definitions, with λ-abstractions as defined in the Conceptual Graph formalism used to describe the covert categories and nouns not included in the hierarchy.

We also saw how for certain nouns, the adult dictionary will find an abstract or complicated nominalization to give a genus. In the AHFD, a noun can be put into a relationship to another part of speech, the most frequent case being that a noun is given as a case relation to a verb.

The limited size of the AHFD, which makes it easier to explore for research purposes is not necessarily an impediment to its usage. The AHFD would be useful to NLP applications, as a natural language teaching tool for children, or a machine translation system for children's books. Also, we believe that the AHFD has multiple characteristics which make it perfect for building the core of a more extended LKB, which can grow later from information extracted from other dictionaries or text corpora.

Notes

[1] The authors would like to thank the anonymous referees for their comments and suggestions. This research was supported by the Institute for Robotics and Intelligent Systems.

[2] Copyright ©1994 by Houghton Mifflin Company. Reproduced by permission from THE AMERICAN HERITAGE FIRST DICTIONARY.

References.

Ahlswede, T. and M. Evens 1988. Generating a relational lexicon from a machine-readable dictionary. *International Journal of Lexicography*, 1(3):214–237.

Amsler, R. 1980. *The structure of the Merriam-Webster pocket dictionary.* Technical Report TR-164, University of Texas, Austin.

Barrière, C. and F. Popowich 1996. Concept clustering and knowledge integration from a children's dictionary. In Proc. of the 16th COLING, Copenhagen, Danemark. Accepted for presentation in August 1996.

Byrd, R., N. Calzolari, M. Chodorow, J. Klavans, M. Neff, and O. Rizk 1987. Tools and methods for computational lexicology. *Computational Linguistics*, 13(3–4):219–240.

Calzolari, N. 1984. Detecting patterns in a lexical data base. In *Proc. of the 10th COLING, Stanford, Cal.*, pp. 170–173.

Chodorow, M.S., R.J. Byrd, and G. Heidorn 1985. Extracting semantic hierarchies from a large on-line dictionary. In *23rd Annual Meeting of the Association for Computational Linguistics*, pp. 299–304, Chicago, Ill.

Cruse, D. 1986, *Lexical Semantics*, Cambridge University Press.

Dolan, W., L. Vanderwende, and S. D. Richardson 1993. Automatically deriving structured knowledge bases from on-line dictionaries. In *The First Conference of the Pacific Association for Computational Linguistics*, pp. 5–14, Harbour Center, Campus of SFU, Vancouver.

Haemmerle, O. 1995. *CoGITo: une plate-forme de developpement de logiciels sur les graphes conceptuels.* PhD thesis, Universite Montpellier II, France.

Markowitz, J., T. Ahlswede, and M. Evens 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of ACL-24*, pp. 112–119.

Montemagni, S. and L. Vanderwende 1992. Structural patterns vs. string patterns for extracting semantic information from dictionaries. In *Proc. of the 14th COLING, Nantes, France*, pp. 546–552.

Sowa, J. 1984. *Conceptual Structures in Mind and Machines*, AddisonWesley.

Wilks, Y., D. Fass, C.-M. Guo, J. McDonald, T. Plate, and B. Slator 1993. Providing machine tractable dictionary tools. In J. Pustejovsky (ed.), *Semantics and the lexicon*, chapter 16, pp. 341–401. KAP.