Retrieving Knowledge for a Lexical Database from Proper Noun Entries in Collins English Dictionary

Muhammad Asadur Rahman
Chicago Stock Exchange, Inc.
440 S. LaSalle Street
Chicago, IL 60605
mrahman@chx.com

Martha Evens
Computer Science Department
Illinois Institute of Technology
Chicago, Illinois 60616
evens@iit.edu

Abstract

This paper presents our efforts to retrieve proper name information from a machine readable dictionary. We explain the problem in brief and then we discuss related work that other researchers have attempted in this area. We describe the embedded knowledge that one can find in the *Collins English Dictionary* (CED). Finally we present our ideas about how to extract this knowledge and how we plan to store it in a lexical database of proper names.

Introduction

Machine readable dictionaries are an excellent source of lexical knowledge. Lexical information is stored in them in a very organized fashion. However, dictionaries are meant for human readers and are not suitable for computer analysis unless the information contained in it is processed and organized for a computer.

Since dictionaries are important sources of lexical knowledge we would like to exploit this source effectively. It is thus necessary to relate different kinds of information provided in a dictionary definition with one another.

Knowledge can be found in a more structured manner in a dictionary. The problem is that knowledge is embedded and entangled in flowing text as descriptions which we human beings understand after we develop the skill of reading the dictionary sublanguage.

Besides the definitions of words, dictionaries contain a wealth of information and knowledge. Since the information is organized and structured we would like to extract the knowledge and store it in a database so that it can be easily retrieved and utilized by a natural language processing program.

We would like to automatically construct the database from a machine readable dictionary so that automatic text generation, information retrieval, text understanding, and intelligent tutoring systems can be supported. Since the dictionary follows a pattern in defining the entries it is possible to identify each type of entry. Once we have identified the entry category, we can figure out what information to look for in the entry and what relations are likely to be hidden in the dictionary text. Our goal is to store the words or phrases under a type category and then store the relationships between them. Our lexical database will store both kinds of information.

Related Research

Gomez and Segami (1994) described a computational model for acquisition of knowledge from encyclopedic text. Their model consisted of three components, namely semantic interpretation, inference, and knowledge representation. This model was implemented into a system called SNOWY which performed complete parsing, interpretation, concept formation, concept recognition and integration in long-term memory. The system was developed to acquire new concepts and conceptual relation about topics dealing with the food habits of animals, their classification and habitats.

Klavans and Chodorow (1990) noted that an independent knowledgebase can be developed by understanding the semantic structure of head nouns. These head nouns such as *unit*, *group*, *number*, and *any*, which are frequently used in dictionary definition provide information other than strict IS-A relationship with the head-word.

Hoang, Strutz, and Evens (1993) identified a few defining strategies and some of the lexical semantic relations used in the *Collins English Dictionary*. They worked with a small subset of proper nouns and manually identified the

defining formulas and the semantic relations between the definition and the proper noun entry. Although the number of proper names used in their study was very small it nevertheless was pioneering work in this area of research. However, they merely scratched the surface of the problem. We are attempting to automatically identify the semantic relations between not only the definition and the entry but also with other entries and the senses within an entry as well.

Conlon et al. (1993) showed that it was possible to construct a lexical database by extracting information from multiple machine readable sources. They outlined the design and construction of their lexical database that was implemented on an Oracle Database management system. They created tables for nouns, verbs, adjectives, and adverbs. There were, however, no separate table for proper names and proper nouns were largly absent from their analysis. Unlike other noun entries, the proper noun entries contain information that identifies and characterizes the proper name. They therefore require extensive analysis and thorough parsing of the definition text in order to extract detailed knowledge for the lexical database.

Calzolari and Picchi (1988) described an approach for word-sense acquisition and knowledge organization from the information contained in a computerized dictionary. Their approach included a discovery procedure technique which was applied recursively and refined on dictionary definitions. They analyzed the free form definition text and converted them into informationally equivalent structures. Their research used the Italian Machine Dictionary (DMI) in lexical database form.

Proper Nouns in CED

Unlike most dictionaries the *Collins English Dictionary* is full of information about proper nouns. The machine readable version of the dictionary has a total of 84,546 entries. About 18,702 or over 22% of them are proper noun entries (Rahman and Evens, 2000). Many proper noun entries have multiple senses and each sense has its own definition text. There are a total of about 24,918 definitions representing different proper noun senses. For proper nouns that are used as plurals the entry in CED contains that information. Proper adjective that are derived from a proper noun entry are listed as a separate sense within the same entry. Generally the entries about people, places, languages, and events contain encyclopedic information and the entries are often large. The size of the definition text for each type of proper names can vary from just one sentence to ten or more sentences. The length of a sentance can be as larg as upto fifty words. Often the sentences are very complex and non-standard in form, which poses a particular challenge in parsing. One has to develop a robust parser to parse and extract all the information. In the following section we will analyze the definition text for different types of proper name entries.

Analysis of Proper Noun Definitions

We have extracted all proper noun entries from the machine readable version of the *Collins English Dictionary*. We are now analyzing the entries to find out how the definition text is constructed, what explicit and implicit information the entries contain and what knowledge is embedded in the definition.

A proper noun definition like other entries in CED has a set of definition text. The first sentence generally defines the proper noun using a special phrasal pattern which other researchers have called *defining formulas.*

```
[1] Aarau : a town in N Switzerland,
    capital of Aargau canton: capital
    of the Helvetic Republic from 1798
    to 1803. Pop.: 16881 (1970).
```

Other sentences generally add more information to the definition.

Different types of proper nouns have different types of key words or phrases that are used as the defining formula. In case of names of individuals the keyword is the occupation of the person such as novelist, poet, engineer, president, etc., which is often preceded by the persons nationality.

```
[2] Milton:  ... 1608-74, English poet.
    His early works ...
```

The headword to definition-head relations are often explicit and can be easily derived from the defining formula. Based on the type of proper noun these relations are can be kinship, spatial, temporal, synonymy, etc. In addition, most entries include IS-A or PART-WHOLE information.

In example [1] above, we can see that there are several semantic relations that are implied in the definition text. First, the head word entry `Aarau` can be described by an IS-A relation with the head noun word `town` in the definition text and it can also be described with a PART-OF relation with `Switzerland` as well as CAPITAL-OF with `Aargau canton`. Note here that the phrase `Aargau canton` leads to a discovery of the fact that `Aargau` is a canton (or province) in `N Switzerland`. Similarly, `Aarau` can be expressed as CAPITAL-OF with `Helvetic Republic`. There are two temporal qualifiers that are associated with this relation: FROM-YEAR and TO-YEAR to denote that it was some time in the past.

Let us consider another example:

```
[3] Pashto:

    (Sense 1):  a  language  of
    Afghanistan  and  NW  Pakistan,
    belonging  to  the  East  Iranian
    branch of the Indo-European family:
    since 1936 the official language of
    Afghanistan.
```

(Sense 2): a speaker of the Pashto language; a Pathan.

(Sense 3): denoting or relating to this language or a speaker of it.

In this example, the relation between the language Pashto and the countries Afghanistan and Pakistan can be expressed with a LANGUAGE-OF relation. We can specify its relation to Pakistan with an adjacency qualifier NORTH-WEST. We can further express Pashto in terms of a taxonomy relation: Pashto IS-A East Iranian Language and East Iranian Language IS-A Indo-European Language. Another relation in sense 1 between Pashto and Afghanistan is OFFICIAL-LANGUAGE-OF. Sense 2 carries only one relation, which is Pashto SPEAKER-OF Pashto Language. The relation implied in this entry is that Pashto IS-A Language.

The entries for names of individuals generally contain more semantic relations. One unique relation that can be found in such entries are the kinship relations. Consider the following definition for Bernoulli:

[4] Bernoulli:

(Sense 1): Daniel. son of Jean Bernoulli. 1700-82, Swiss mathematician and physicist, who developed an early form of the kinetic theory of gases and stated the principle of conservation of energy in fluid dynamics.

(Sense 2): Jacques. 1654-1705, Swiss mathematician, noted for his work on calculus and the theory of probability.

(Sense 3): his brother, Jean. 1667-1748, Swiss mathematician who developed the calculus of variations.

By studying the definition text we can find information about a persons' name, gender, when born and when died (if applicable), where born, nationality, profession, and what the person is famous for. In case of person names the CED in general provides information about most of these items. Furthermore, we can derive the following lexical-semantic relations from the above proper name entry.

Daniel Bernoulli	PROFESSION	Mathematician
Daniel Bernoulli	PROFESSION	Physicist
Daniel Bernoulli	DEVELOPED	Kinetic Theory of Gases
Daniel Bernoulli	STATED	Principle of Conserv of Energy
Daniel Bernoulli	SON-OF	Jean Bernoulli
Jacques Bernoulli	PROFESSION	Mathematician
Jacques Bernoulli	NOTED-FOR	Work on Calculus
Jacques Bernoulli	NOTED-FOR	Work on Calculus
Jacques Bernoulli	NOTED-FOR	Theory of Probability
Jean Bernoulli	BROTHER-OF	Jacques Bernoulli
Jean Bernoulli	PROFESSION	Mathematician
Jean Bernoulli	DEVELOPED	Calculus of Variations

The construction and information content in proper noun definitions are very diverse in general and even within the same type of proper names they provide different kinds of information. However, there are still some common features, and if we look very closely we will find some patterns. Our plan is to use these common types of information or the patterns as fields in a table or from the object oriented point of view as attributes of a general class of proper names.

Kinship relations used in CED definitions of proper names:

Brother of	*father-in-law of*
Sister of	*mother-in-law of*
Father of	*uncle of*
Mother of	*cousin of*

Spatial relations used in CED definitions of proper names are *from, between, near, within* etc. Synonymy relations used in CED definitions of proper names are:

Former name of	*also called*
Abbreviation for	*also known as*
acronym for	*Symbol for*
another name for	*trade mark for*

Part-whole relations used in proper name definitions are *member of, belongs to, belonging to, group of*, etc.

We have extracted all proper noun entries from the machine readable version of the *Collins English Dictionary*. We are now analyzing the entries to find out how the definition text is constructed, what explicit and implicit information they contain and what knowledge is embedded in the definition.

While defining the template we have considered the kind of information the definition text provides in the dictionary for each type of proper name. In fact we envision each category and sub-category of proper names as objects and each kind of information as attributes of that object. Thus for the entries of individuals we have first name, last name, nationality, year of birth, year of death (if the person is not alive), profession, kinship with other famous personalities, etc. This has helped us in designing the parsing process and development effort for the parser we are using in parsing the definition texts. We will be extracting the above mentioned information along with the relations into an database.

Strategies for Extracting Knowledge

In the previous section we have presented the analysis of a few examples of proper noun definitions from the *Collins English Dictionary*. We analyzed them to reveal the kinds of explicit information and knowledge they contain. We have also indentified all explicit and implicit semantic relations those definitions contained. In this section we would like to outline the strategies we are adopting in order to identify and extract all information and semantic relations automatically from all the definition texts.

The first step in our strategy involves the classification of proper names. We have already developed a proper name classification scheme based on the types of proper noun entries found in the CED (see Rahman and Evens, 2000). The next step is to identify the defining formula in each definition text. This was also done when we were developing the classification scheme. Besides identifying the types of proper name, this step also revealed the semantic relation the headword entry has with the head noun in the definition text. At this point we have identified only one semantic relation per definition text. Here, we employed a simple parser with a few pattern-matching rules that extracted the defining formula and subsequently found the primary semantic relation by identifying the head noun. For example, when parsing the definition text for `Kauai`:

```
[5] Kauai: a volcanic island in NW
    Hawaii, ...
```

The parser finds the defining formula as *a volcanic island in* and the head noun *island*. So, it determines the primary semantic relation of the definition to be:

```
Kauai      IS-A    island
```

In order to find the defining formula the parser looked for the end of sentence for the first sentence in the definition text and discarded the rest of the definition text. There are exception, however, in case of some entries for names of individuals, the first sentence contains the first name which follows the year of birth and if applicable, the year of death. Here is such an example:

```
[6] Chichester: Sir Francis.  1901-72,
    English yachtsman, who sailed ...
```

The parser starts to scan the sentence from left to right identifying each word. The first word in example `[5]` is *a* which is an article and the parser recognizes it by looking up a list of English articles. The next word it encounters is *volcanic* which it recognizes to be an adjective by looking up the word in the CED. The next word *island* is recognized as a noun through yet another dictionary lookup. The parser flags the nouns as possible candidates for the head noun in the definition. The next word the parser finds is *in* which is recognized as a preposition. Since this is a terminal word for a defining formula the parsing process stops. The parser identified the word to be a terminal by matching it against a list of terminal words. If the parser finds more than one noun before it reaches the terminal word it must resolve the potential ambiguity in finding the head noun. In example `[6]` the parser determines that *Sir* is a title which followed by the first name. The year of birth is *1901* and the year of death is *1972* . It finds that the word *English* could both be a noun and an adjective, and the word *yachtsman* is recognized to be a noun through a dictionary lookup. One of our pattern matching rule says that a word which precedes a noun and could be either a noun or an adjective, it is most likely to be an adjective. Thus the word *yachstman* is recognized as a head noun. The parser suggests the following semantic relation:

```
Sir Francis Chichester  IS-A yachtsman
```

The technique our parser used so far is pure pattern matching with some smart rules. This is sufficient to identify the defining formula and the head noun typically from the first or the second sentence of a dictionary definition. In order to parse the rest of the definition text a more sophisticated natural language parser is needed. The definition text however still contains information that can be extracted by employing further pattern matching. Our strategy at this point is to enhance the current parser with more pattern matching rules to extract the information that does not require sophisticated parsing. For example, names of individuals have year of birth, year of death (if applicable), nationality information, cities have population information along with the year of census or year of estimation, countries have information about the area, the name of the capital, besides the population information. Similarly, mountains and rivers have information about height and length respectively. Once we are able to extract the above information we will upgrade our parser to be a full blown NLP parser to parse all other sentences from the definition text. Until this is done we will have one semantic relation per definition.

An important aspect of our work is to organize and store the extracted knowledge and information in a lexical database of proper nouns. The outline for a lexical database discussed by Conlon and Evens (1994) is not suitable for storing the kind of information we are extracting. We already defined the tables for all different categories of proper names according to out proper name classification scheme. However, as we continue parsing the definition text we may find that we need to add more fields to a certain table or some fields may be considered redundant or unnecessary. So, we are flexible about our design.

Conclusions

We have demonstrated that the knowledge contents in a machine readable dictionary can be effectively extracted and stored in a lexical database by establishing the relations between the individual words and phrases and storing those relations. Our approach is to analyze the definition text to identify the explicit and implicit information content and the semantic relations that exist in the text. We will then convert the entire definition text into informationally equivalent structure of knowledge and semantic relations. We have grouped the word-senses for similar classes of proper names and by studying the definition texts for all classes we have identified the type of information that the definition text contains.

Acknowledgements

The authors would like to thank Collins Publishers and Patrick Hanks for their permission to use the *Collins English Dictionary* for his research. Thanks also to the Data Collection Initiatives (DCI) for providing us with the machine readable version of the dictionary.

References

Ahlswede, Thomas. 1988. Syntactic and Semantic Analysis of Definitions in a Machine-Readable Dictionary, Ph.D. Thesis, Department of Computer Science, Illinois Institute of Technology, Chicago, IL.

Amsler, Robert. 1981. A Taxonomy for English Nouns and Verbs. Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics. 133-138. Stanford, CA.

Atkins, B. T., Kegl J., and Levin B. 1986. Explicit and Implicit Information in Dictionaries. In Proceedings of the Second Annual Conference of the UW Center for the New Oxford English Dictionary. 45-63. Waterloo, Ontario, Canada: UW Center for the New OED and Text Research.

Boguraev, B., and Levin, B. 1993. Models for Lexical Knowledge Base. J. Pustejovsky (ed.), Semantics and the Lexicon, ' 325-340. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Calzolari, N. and Picchi, E. 1988. Acquisition of Semantic Information from an On-line Dictionary. In Proceedings of the COLING-88. 87-92. Budapest, Hungary.

Coates-Stephens, S. 1991. Automatic Lexical Acquisition Using Within-Text Descriptions of Proper Nouns, In Proceedings Seventh Annual Conference of the University of Waterloo (UW) Center for New Oxford English Dictionary (OED) and Text Research, 154-169. Waterloo, Canada: UW Center for the New OED and Text Research.

Coates-Stephens, S. 1993. The Analysis and Acquisition of Proper Names for the Understanding of Free Text, Computers and the Humanities 26(4):441-456.

Conlon, S. P., and Evens, M. W. 1994. A Lexical Database for Nouns to Support Parsing, Text Generation, and Information Retrieval, In S. Hockey and N. Ide, eds., Research in Humanities Computing 3. 74-87. Oxford: Clarendon Press.

Conlon, S. P., Evens, M., Ahldwede, T. and Strutz, R. 1993. Developing A Large Lexical Database for Information Retrieval, Parsing, and Text Generation Systems. Information Processing & management 29(4): 415-431.

Gomez, F., Hull, R., and Segami, C. 1994. Acquiring Knowledge from Encyclopedic Text. In Proceedings of the Fourth Conference on Applied Natural Language Processing, 84-90. Stuttgart, Germany: Association for Computational Linguistics.

Hanks P. (ed.). 1979. Collins English Dictionary, First Edition, William Collins Sons & Co. Ltd., London, England. [Machine Readable Version. Data Collection Initiative (DCI).]

Hoang, H. L., Strutz, R., and Evens, M. 1993. Lexical Semantic-Relations for Proper Nouns. In Proceedings of Fourth Annual Midwest Artificial Intelligence and Cognitive Science (MAICS) Conference. 26-30. Cheslerton, Indiana: AAAI Press.

Kim J. 1996. Extracting Lexical Database Entries for Proper Nouns from the Wall Street Journal. Ph.D. Thesis, Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois.

Klavans, J. L., and Chodorow, M. S. 1990. From Dictionary to Knowledge Base via Taxonomy. In Proceedings of the Sixth Annual Conference of the UW Center for the New Oxford English Dictionary and Text Research. 110-132. Waterloo, Ontario, Canada: UW Center for the New OED and Text Research.

Markowitz, J., Ahlswede, T., and Evens, M. 1986. Semantically Significant Patterns in Dictionary Definitions. In Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics. 112-119. New York, New York: AAAI Press.

Meijs, W. 1993. Exploring Lexical Knowledge. In Corpus-Based Computational Linguistics. Souter, C. and Atwell E. (eds.). 249-260. Atlanta, GA: Rodopi.

Mufwene, S. S. 1988. Dictionaries and Proper Names. International Journal of Lexicography 1(3): 268-283.

Nutter, J. T., Fox, E. A., and Evens, M. W. 1994. Building a Lexicon from Machine-Readable Dictionaries for Improved Information Retrieval. Literary and Linguistic Computing 2(5):1-77.

Rahman, M. A. and Evens M. 2000. *Categorization of Proper Names for Inclusion in a Lexical Database.* Proc. of the International Conference on Computers and their Applications (CATA-2000), New Orleans, FL.

Tompa, F. W. and Raymond, D. R. 1989. Database Design for a Dynamic Dictionary, Technical Report, OED-89-05, UW Center for the New Oxford English Dictionary, Waterloo, Ontario, Canada.