Language, Information and Computation(PACLIC 11), 1996, 137-146

Extraction of Thematic Roles from Dictionary Definitions

Dr. Michael L. Mc Hale
RL/C3CA
Rome Laboratory
525 Brooks Road
Rome, New York
13441-4505

mchale@ai.rl.af.mil

Prof. Sung H. Myaeng
Department of Computer Science
Chungnam National University
220 Kung-dong, Yoosung-ku
Taejon, 305-764
Korea

shmyaeng@cs.chungnam.ac.kr

Abstract

Our research goal has been the development of a domain independent natural language processing (NLP) system suitable for information retrieval. As part of that research, we have investigated ways to automatically extend the semantics of a lexicon derived from machine-readable lexical sources. This paper details the extraction of thematic roles derived from lexical patterns in a machine-readable dictionary.

Introduction

With information retrieval as a goal, we are very sensitive to the issues of generalization, speed and scalability. Any NLP system that is used for information retrieval must be capable of handling large amounts of general text in a timely manner. Each of the components of such a system, from morphology through semantics, must have similar capabilities. Thematic roles, which provide a very basic "who does what to whom" type of semantics, meet these criteria because they are very general, simple and useable by a wide variety of natural language processing systems. The roles are generally contained in frames that contain the type of each argument for each verb. For example *eat* would have a frame similar to *eat*[AGENT, THEME].

We have explored the development of these frames by using information found in an on-line version of *Longman's Dictionary of Contemporary English* (LDOCE 1987). The information in the dictionary includes: definitions; subject field codes; the box codes, that provide information on the type of arguments (ex., human or abstract); a reduced set of grammar codes, that provide information on the transitivity of a verb and the syntactic category of any extra arguments; and other information much of which is probably extraneous to the roles.

We focused on the definitions because we felt that they would provide the best base for a scaleable approach. Not only do the definitions contain information germane to role extraction but they are complete in the sense that every verb has a definition. This is not the case for the other types of information available. The box codes, for instance, are generic (i.e., empty) codes around 17% of the time.

Our approach to analyzing the definitions is based on lexical patterns. A lexical pattern is simply a series of consecutive words that is used in more than one definition. Some of these have the appearance of a syntactic pattern (ex., *to cause to*) while others more directly reflect their lexical nature (ex., *longer have because*). Lexical patterns are the logical place to start because they are the lowest level of analysis that seems likely to contain sufficient information for role extraction. Obviously, we could have included tagging, syntactic analysis, syntactic patterns, statistics on the box codes or some other information. By focusing on the lowest level of processing, however, we ensure that any information that is extraneous to the process (ex., syntax) is ignored and moreover, that our results will be easily repeatable and scaleable to general text processing.

We used a modified Matrix Model (Cook 1989) as a template for the roles. The Matrix Model (Figure 1) has five roles: **T** (or **Ts**) a Theme, **A** an Agent, **B** a Benefactor, **E** an Experiencer and **L** a

Verb Types	Basic	Experiential	Benefactive	Locative
1. State	Ts *be tall* Ts, Ts *be + N*	E, Ts *like* Ts, E *be boring*	B, Ts *have* Ts, B *belong to*	Ts, L *be in* L, Ts *contain*
2. Process	T *die* T, T *become*	E, T *enjoy* T, E *amuse*	B, T *acquire* T, B ...	T, L *move, in* L, T *leak*
3. Action	A, T *kill* A, T, T *elect*	A, E, T *say* A, T, E *amuse (agt)*	A, B, T *give* A, T, B *blame*	A, T, L *put* A, L, T *fill*

Figure 1. Matrix Model
(Adapted from Cook 1989)

Location. The model is computationally attractive as it allows for classification of thematic roles (Case Grammar) into discrete groups.

This allows the assignment of role frame by determining the proper row and then the proper column for the verb.

Methodology

The overall approach is based on two main assumptions: 1) Cook's matrix model is correct and computationally feasible, and 2) the repetitive nature of the definitions in *LDOCE* provide sufficient lexical clues allowing lexical patterns to be used to extract the thematic roles.

The computational feasibility of the model will be demonstrated below. The correctness of the model is of somewhat more concern. The term correctness is not being used here to denote psychological correctness but rather computational correctness. For the matrix to be computationally correct the divisions (rows and columns) must be both exhaustive and mutually exclusive. That is, each thematic role must be assignable to one and only one square in the matrix. Cook covers these concerns to our satisfaction in his section on the design of the matrix and they therefore will not be covered here.

The use of lexical patterns in definitions has successfully been exploited by a number of different researchers (Ahlswede 1988, Wilkins 1988, McHale 1991, 1995). It was anticipated that the patterns found in the definitions of verbs in *LDOCE* would be useful in discriminating among the rows of the Matrix Model. For instance, the phrase *to cause to* in an *LDOCE* definition generally indicates an action verb. Once the verbs were categorized by matrix row then clues would be sought to differentiate them by column.

Some of the clues could provide positive evidence (ex., *to cause to*) while others provide negative evidence (ex., a box code of *human* eliminates the consideration of a process-locative verb). An earlier version of *LDOCE* contained a much richer set of grammar codes but was inconsistent and incomplete. After receiving numerous complaints about the grammar codes, Longman's decided to eliminate most of them. Thus, the later version, which we are using, has a more consistent but much reduced set. Some researchers (cf. Dorr 93) have used the earlier version to extract thematic roles. However, the techniques thus developed are not generalizable even to the later version of the same dictionary.

The lexical patterns were determined in the following manner. Appendix 1 of Cook gives 320 verb senses with their associated frames. The definition entries in *LDOCE* that correspond to those verbs were extracted. These verb senses were then checked to ensure that the proper sense of the verb was associated with each frame. Once this was done, the verbs, and their associated frames, were

grouped in two ways: once for the row of the matrix in which they would occur and once for the column. For example, all words with AGENT in their frame were put in the action group and all words with LOCATIVE were put in the locative group.

Each group was then analyzed for lexical patterns. This was accomplished by producing each 2-, 3-, 4-, 5-, 6-, 7-, 8- and 9-word group that is present in each definition. For example, *to cause to cry*, would have one 4-word pattern (*to cause to cry*), two 3-word patterns (*to cause to* and *cause to cry*) and three 2-word patterns (*to cause*, *cause to*, and *to cry*). Note that all these patterns must be tested. Obviously, if a definition contains *to cause* then it can be used to find *to cause to cry*. Thus, it might be assumed that only the shorter, two-word pattern need be maintained. However, the shorter pattern may occur in a variety of frames and therefore provide weaker discriminatory power. There is no way of discovering the discriminatory power of each pattern without initially testing and maintaining all the patterns.

The patterns for each row (STATE, PROCESS and ACTION) were maintained separately. After all the patterns were extracted for each row, the separate groups were sorted and all duplicate patterns were eliminated from them. Then the patterns for the three rows were combined, re-sorted and those patterns that occurred in more than one row were eliminated from the separated groups. This left in each separated group only the patterns that were unique to each associated row. This resulted in 38,335 unique 2-9 word patterns: 5,014 STATIVE; 6,692 PROCESS; and 26,629 ACTION.

Extraction of Frames

The unique patterns were then used to process the definitions of the whole dictionary. Each of the 11,931 verb sense definitions was processed to determine if it contained one of the 38,335 unique patterns. Those definitions that contained a pattern were considered as potentially having a frame that belonged to the row to which the pattern belonged (ex., *to cause to* - ACTION).

The results of the extraction process are shown in Table 1. The first column is the number of words in the pattern. Columns 2-4 are the number of definitions associated with the ACTION, PROCESS and STATIVE rows respectively. (For example, there were 7194 verb senses classified as action verbs through the use of 2-word lexical patterns.) Column 5 (total) is the sum of columns 2-4 and represents the total number of verb senses classified. Column 6 (unique) is column 5 with the duplicates removed. Column 7 (overlap) presents

the number of definitions associated with two or more rows, thus it is calculated as column 5 minus column 6.

1	2	3	4	5	6	7
Words	ACTION	PROCESS	STATIVE	Total	Unique	Overlap
2	7194	3970	2914	14078	8900	5178
3	5112	1041	979	7132	6134	998
4	2614	452	911	3977	3353	624
5	1242	261	246	1749	1717	32
6	805	163	151	1119	1119	0
7	511	122	105	738	738	0
8	421	105	83	609	609	0
9	367	96	64	527	527	0

Table 1. Details from Row Extraction

Table 2 shows a variety of ways of combining these data by the length of the patterns. It also shows the percentages of extraction and overlap resulting from each combination. The first column is the size of the patterns used. The first row, for instance, uses all the patterns of length 2 through length 9. Columns 2-4 are the number of definitions extracted for ACTION, PROCESS and STATIVE verbs respectively using the given combination. Column 5 is the percentage of definitions extracted. This column is computed by adding columns 2-4 and dividing by the total number of verb definitions. This value can exceed 100% because there may be overlaps in assigning roles. Column 6 is the percentage of overlap produced. It is computed by taking the difference between the sum of columns 2-4 and the sum of columns 2-4 not counting duplicates. The difference is then divided by the total number of verb definitions to produce the percentage of overlap.

1	2	3	4	5	6
Patterns	ACTION	PROCESS	STATIVE	Extracted	Overlap
2-9	7795	4207	3596	131%	53%
3-9	5176	1091	1439	65%	13%
4-9	2624	467	916	34%	5%
5-9	1248	261	247	15%	0.3%
6-9	806	163	151	9%	0%

Table 2. Combination of Rows

Table 2 shows the trade-off between the degree of extraction and the amount of overlap. The shorter patterns produced classifications for most of the verbs but could not do so uniquely. For instance,

using all the patterns of length 2 to length 9 produces 15598 role extractions for 9261 (78%) of the 11931 verb definitions. Many of the definitions are given two or three frames accounting for the 53% overlap. The longer patterns (6-9) produced only unique classifications but could do so for only 1120 (9%) of the verbs. The best balance between the amount of verbs classified and the percentage of role plurality seems to be 3-9 with 65% classification and 13% plurality. This was a strong result but still left some room for improvement. The goal was to maximize the number of frames extracted while minimizing the amount of overlap. We explored two ways to approach this.

Enhancement Techniques

The first technique would use the combination with the most extractions (2-9) and attempt to minimize the overlap by using other information available from the dictionary (ex., box codes, subject field codes). Attempts to do this by hand have been less than encouraging. No consistent methodology to reduce the overlap that was not based either on world knowledge or on an *ad hoc* method has been found. Therefore this approach was abandoned.

The second approach would use the combination with no overlap (6-9) to *bootstrap* the rest of the patterns. That is, if the 1120 verb senses extracted by the patterns have been correctly categorized then it should be possible to analyze them for new, unique patterns that can then be used to find more verb senses of the same type. This approach relies on the validity of those frames already extracted. Thus, the degree of correctness of the frames had to be verified before this approach could be used. To that end, a random sample from the 1120 verb senses was taken and their respective roles were determined by hand. These roles were then compared to the algorithm output. The result was that 89% of the extractions were correct. We felt that this was sufficiently precise to warrant further investigation of the "bootstrap" approach.

The 1120 extracted verb senses were subsequently analyzed for lexical patterns. Again, all the 2-, 3-, 4-, 5-, 6-, 7-, 8- and 9-word groups in each definition were produced. This resulted in 60,122 patterns of which 56,627 were unique. These patterns were again used to do the extraction from the whole dictionary. The results of this extraction are given in Tables 3 and 4.

1	2	3	4	5	6	7
Words	ACTION	PROCESS	STATIVE	Total	Unique	Overlap
2	8329	2080	1611	12020	9082	2938
3	5604	1039	1017	7660	6535	1125
4	2874	439	442	3755	3561	194
5	1412	252	241	1905	1876	29
6	890	166	184	1240	1239	1
7	777	149	153	1079	1079	0
8	703	136	114	953	953	0
9	642	126	97	865	865	0

Table 3. Re-extraction of Rows

1	2	3	4	5	6
Pattern	ACTION	PROCESS	STATIVE	Extracted	Overlap
2-9	8560	2449	1995	109%	31%
3-9	5654	1091	1075	66%	10%
4-9	2882	446	458	32%	2%
5-9	1413	252	241	16%	0%
6-9	891	166	184	10%	0%

Table 4. Re-Combination of Rows

These two tables represent a significant amount of processing yet there is almost no change in the overall result; Table 4 looks remarkably similar to Table 2. In fact, for all the processing that was required there were only twenty-one verb senses categorized through the re-extraction that were not originally categorized by the algorithm. It should be obvious to the most casual observer that this minute improvement in extraction cannot justify the tremendous amount of processing required to produce it. Therefore we cannot justify using the bootstrap method of extraction enhancement and have abandoned it also.

The bottom line for row extraction then is 65% extraction with 13% overlap using lexical patterns in this manner. These results may have been a consequence of the interaction between the *LDOCE* defining vocabulary and the row designators of the matrix. To ensure that was not the case we repeated the whole process with the columns.

Extraction of Columns

The results are slightly less encouraging than that for the rows. There were 38,634 unique patterns extracted: 10,468 BASIC; 6,930 BENEFACTIVE; 5,390 EXPERIENTIAL; and 15,846 LOCATIVE. These patterns were found in 9,154 definitions (77%) with 62% overlap.

BASIC and LOCATIVE created the most overlap, but all the columns contributed. Tables 5 and 6 give the results. The computations are done in the same way that they were for Tables 1 through 4.

1	2	3	4	5	6	7	8
Words	BASIC	BENE	EXP	LOC	Total	Unique	Overlap
2	9887	9446	9179	10079	38591	10514	28077
3	6273	4892	4700	6723	22588	7835	14753
4	2645	1957	1261	3058	8921	4487	4434
5	1194	907	465	1451	4017	2415	1602
6	631	246	193	763	1833	1424	409
7	321	155	122	431	1029	906	123
8	261	127	96	352	836	741	95
9	226	112	79	300	717	639	78

Table 5. Extraction of Columns

1	2	3	4	5	6	7
Pattern	BASIC	BENE	EXP	LOC	Extracted	Overlap
2-9	9887	9446	9179	10079	323%	235%
3-9	6273	4892	4700	6723	189%	124%
4-9	2645	1957	1261	3058	75%	37%
5-9	1194	907	465	1451	34%	13%
6-9	631	246	193	763	15%	3%

Table 6. Combination of Columns

The results for the columns indicate that the general problem may be one of overlap and not extraction. The best result is perhaps the 4-9 combination which yields 75% extraction but with 37% overlap. What causes the overlap? The cause can be shown with the definition of the verb *enlighten*. *LDOCE* defines it as:

enlighten - to cause to understand deeply and clearly, esp. by making free from false beliefs.

The definition contains both the agentive pattern *to cause to* and the stative pattern *to understand*. While this particular combination (to cause to + stative) is rare in *LDOCE* (occurring with only four other verbs) it is this conflict of double patterns that causes the overlap in all cases. The presence of double patterns appears to be, in part, a result of the restricted defining vocabulary used in *LDOCE*. The limited vocabulary creates the abundance of lexical patterns that make the approach possible but the vocabulary is so limited that the patterns cannot be uniquely used for a given type of verb.

We approached this task assuming also that the definitions used in the dictionary contained sufficient patterns to facilitate the extraction of thematic roles. The results lend credence to that assumption. The patterns do facilitate the extraction but they are not sufficient by themselves. The results indicate that the best we can hope for is around 65-70% extraction with 10-15% overlap and 90% accuracy. This is not sufficient for a totally automated system but should be a solid basis for a semi-automatic tool to assist in determining thematic roles. The creation of such a mixed initiative extraction system is a logical next step for our research. The system would do the extraction analysis and assign roles for those verb senses where it could do so unambiguously. For the rest, it could present the results to the user along with all other pertinent information (the definition, box codes, example sentences, etc.).

Summary and Discussion

This paper examines the extraction of thematic roles from dictionary definitions. The approach is based on lexical patterns (word co-locations) and not on syntactic structure. The reasons for this choice were both pragmatic and theoretic. Pragmatic in that a syntactic approach would be much more complicated. It is relatively easy to find all occurrences of *to cause to* but a syntactic approach would probably have to consider *to cause (VP)* or perhaps *(VP) (VP)* or some other combination. The number of combinations using lexico-syntactic information is therefore much larger than straight lexical patterns. The choice was theoretic in that the use of lexico-syntactic information must be considered overkill until the efficacy of straight lexical patterns was examined.

What this research shows is not that the lexical patterns alone do not work but that they do not work well enough to be used for totally automatic extraction. In general, the results were around 65% extraction with 10-15% overlap and 90% accuracy. Efforts to increase the precision of the extraction proved fruitless leaving us with the realization that this approach would best be used as a firm foundation for creating a mixed initiative (human-computer) extraction system.

The reason for these results appears to be in part the nature of the definitions in *LDOCE*. By confining the definitions to a very small defining vocabulary many of the phrases (i.e., lexical patterns) have to do double duty and are therefore used in definitions of words with different thematic roles. Further research should be carried out using a different dictionary to see if a richer defining vocabulary still has sufficient patterns to allow our method to work. The other area

of further research is in the use of syntactic patterns. It is our opinion that the latter promises to be a much more fruitful area of research.

References

Ahlswede, T.E. 1988. *Syntactic and Semantic Analysis of Definitions in a Machine-Readable Dictionary*, Ph.D. Dissertation, Illinois Institute of Technology.

Cook, W. 1989. *Case Grammar Theory*. Georgetown University Press: Washington.

Dorr, B. 1993. *Machine Translation: A View from the Lexicon*. Artificial Intelligence Series. The MIT Press, Cambridge, Mass.

LDOCE 1987. *Longman Dictionary of Contemporary English*. Longman: Harlow, UK.

Mc Hale, M. L. 1991. *The Production of a Parser for Longman's Dictionary of Contemporary English*. IEEE Dual-Use Technology Conference, State University of New York, Utica, New York. June 1991.

Mc Hale, M. L. 1995. *Combining Machine-Readable Lexical Resources with a Principle-Based Parser*, Ph.D. Dissertation, Syracuse University, NY.

Wilkins, E. 1988. *Syntax and Semantics: Thematic Relations,* Academic Press, Inc.: San Diego, CA.