Combining Multiple, Large-Scale Resources in a Reusable Lexicon for Natural Language Generation

Hongyan Jing and **Kathleen McKeown**
Department of Computer Science
Columbia University
New York, NY 10027, USA
{hjing, kathy}@cs.columbia.edu

Abstract

A lexicon is an essential component in a generation system but few efforts have been made to build a rich, large-scale lexicon and make it reusable for different generation applications. In this paper, we describe our work to build such a lexicon by combining multiple, heterogeneous linguistic resources which have been developed for other purposes. Novel transformation and integration of resources is required to reuse them for generation. We also applied the lexicon to the lexical choice and realization component of a practical generation application by using a multi-level feedback architecture. The integration of the lexicon and the architecture is able to effectively improve the system paraphrasing power, minimize the chance of grammatical errors, and simplify the development process substantially.

1 Introduction

Every generation system needs a lexicon, and in almost every case, it is acquired anew. Few efforts in building a rich, large-scale, and reusable generation lexicon have been presented in literature. Most generation systems are still supported by a small system lexicon, with limited entries and hand-coded knowledge. Although such lexicons are reported to be sufficient for the specific domain in which a generation system works, there are some obvious deficiencies: (1) Hand-coding is time and labor intensive, and introduction of errors is likely. (2) Even though some knowledge, such as syntactic structures for a verb, is domain-independent, often it is re-encoded each time a new application is under development. (3) Hand-coding seriously restricts the scale and expressive power of generation systems. As natural language generation is used in more ambitious applications, this situation calls for an improvement.

Generally, existing linguistic resources are not suitable to use for generation directly. First, most large-scale linguistic resources so far were built for language interpretation applications. They are indexed by *words*, whereas, an ideal generation lexicon should be indexed by the *semantic concepts* to be conveyed, because the input of a generation system is at semantic level and the processing during generation is based on semantic concepts, and because the mapping in the generation process is from concepts to words. Second, the knowledge needed for generation exists in a number of different resources, with each resource containing a particular type of information; they can not currently be used simultaneously in a system.

In this paper, we present work in building a rich, large-scale, and reusable lexicon for generation by combining multiple, heterogeneous linguistic resources. The resulting lexicon contains syntactic, semantic, and lexical knowledge, indexed by senses of words as required by generation, including:

- A complete list of syntactic subcategorizations for each sense of a verb to support surface realization.
- A large variety of transitivity alternations for each sense of a verb to support paraphrasing.
- Frequency of lexical items and verb subcategorizations and also selectional constraints derived from a corpus to support lexical choice.
- Rich lexical relations between lexical concepts, including hyponymy, antonymy, and so on, to support lexical choice.

The construction of the lexicon is semi-automatic, and the lexicon has been used for lexical choice and realization in a practical generation system. In Section 2, we describe the process to build the generation lexicon by combining existing linguistic resources. In Section 3, we show the application of the lexicon by actually using it in a generation system. Finally, we present conclusions and future work.

2 Constructing a generation lexicon by merging linguistic resources

2.1 Linguistic resources

In our selection of resources, we aim primarily for accuracy of the resource, large coverage, and providing a particular type of information especially useful for natural language generation. four linguistic resources:

1. The WordNet on-line lexical database (Miller et al., 1990). WordNet is a well known on-line dictionary, consisting of 121,962 unique words, 99,642 synsets (each synset is a lexical concept represented by a set of synonymous words), and 173,941 senses of words.[1] It is especially useful for generation because it is based on *lexical concepts*, rather than words, and because it provides several semantic relationships (hyponymy, antonymy, meronymy, entailment) which are beneficial to lexical choice.
2. English Verb Classes and Alternations (EVCA) (Levin, 1993). EVCA is an extensive linguistic study of diathesis alternations, which are variations in the realization of verb arguments. For example, the alternation "there-insertion" transforms *A ship appeared on the horizon* to *There appeared a ship on the horizon*. Knowledge of alternations facilitates the generation of paraphrases. (Levin, 1993) studies 80 alternations.
3. The COMLEX syntax dictionary (Grishman et al., 1994). COMLEX contains syntactic information for 38,000 English words. The information includes subcategorization and complement restrictions.
4. The Brown Corpus tagged with WordNet senses (Miller et al., 1993). The original Brown corpus (Kučera and Francis, 1967) has been used as a reference corpus in many computational applications. Part of Brown Corpus has been tagged with WordNet senses manually by the WordNet group. We use this corpus for frequency measurements and exacting selectional constraints.

[1] As of Version 1.6, released in December 1997.

2.2 Combining linguistic resources

In this section, we present an algorithm for merging data from the four resources in a manner that achieves high accuracy and completeness. We focus on verbs, which play the most important role in deciding phrase and sentence structure.

Our algorithm first merges COMLEX and EVCA, producing a list of syntactic subcategorizations and alternations for *each verb*. Distinctions in these syntactic restrictions according to each *sense* of a verb are achieved in the second stage, where WordNet is merged with the result of the first step. Finally, the corpus information is added, complementing the static resources with actual usage counts for each syntactic pattern. This allows us to detect rarely used constructs that should be avoided during generation, and possibly to identify alternatives that are not included in the lexical databases.

2.2.1 Merging COMLEX and EVCA

Alternations involve syntactic transformations of verb arguments. They are thus a means to alleviate the usual lack of alternative ways to express the same concept in current generation systems.

EVCA has been designed for use by humans, not computers. We need therefore to convert the information present in Levin's book (Levin, 1993) to a format that can be automatically analyzed. We extracted the relevant information for each verb using the verb classes to which the various verbs are assigned; members of the same class have the same syntactic behavior in terms of allowable alternations. EVCA specifies a mapping between words and word classes, associating each class with alternations and with subcategorization frames. Using the mapping from word and word classes, and from word classes to alternations, alternations for each verb are extracted.

We manually formatted the alternate patterns in each alternation in COMLEX format.

The reason to choose manual formatting rather than automating the process is to guarantee the reliability of the result. In terms of time, manual formatting process is no more expensive than automation since the total number of alternations is small(80). When an alternate pattern can not be represented by the labels in COMLEX, we need to added new labels during the formatting process; this also makes automating the process difficult.

The formatted EVCA consists of sets of applicable alternations and subcategorizations for 3,104 verbs. We show the sample entry for the verb *appear* in Figure 1. Each verb has 1.9 alternations and 2.4 subcategorizations on average. The maximum number of alternations (13) is realized for the verb "roll".

The merging of COMLEX and EVCA is achieved by unification, which is possible due to the usage of similar representations. Two points are worth to mention: (a) When a more general form is unified with a specific one, the later is adopted in final result. For example, the unification of PP[2] and PP-PRED-RS[3] is PP-PRED-RS. (b) Alternations are validated by the subcategorization information. An alternation is applicable only if both alternate patterns are applicable.

Applying this algorithm to our lexical resources, we obtain rich subcategorization and alternation information for each verb. COMLEX provides most subcategorizations, while EVCA provides certain rare usages of a verb which might be missing from COMLEX. Conversely, the alternations in EVCA are validated by the subcategorizations in COMLEX. The merging operation produces entries for 5,920 verbs out of 5,583 in COMLEX and 3,104 in EVCA.[4] Each of these verbs is associated with 5.2 subcategorizations and 1.0 alternation on average. Figure 2 is an updated version of Figure 1 after this merging operation.

```
appear:
((INTRANS)
 (LOCPP)
 (PP)
 (ADJ-PER-PART)
 (INTRANS THERE-V-SUBJ :ALT There-Insertion)
 (LOCPP THERE-V-SUBJ-LOCPP :ALT There-Insertion)
 (LOCPP LOCPP-V-SUBJ :ALT Locative_Inversion))
```

Figure 1: Alternations and subcategorizations from EVCA for the verb *appear*.

```
appear:
((PP-TO-INF-RS :PVAL ("to"))
 (PP-PRED-RS :PVAL ("to" "of" "under" "against"
             "in favor of'' "before" "at"))
 (EXTRAP-TO-NP-S)
 (INTRANS)
 ...
 (INTRANS THERE-V-SUBJ :ALT There-Insertion)
 (LOCPP THERE-V-SUBJ-LOCPP :ALT There-Insertion)
 (LOCPP LOCPP-V-SUBJ :ALT Locative_Inversion)))
```

Figure 2: Entry for the verb *appear* after merging COMLEX with EVCA.

2.2.2 Merging COMLEX/EVCA with WordNet

WordNet is a valuable resource for generation because most importantly the synsets provide a mapping between concepts and words. Its inclusion of rich lexical relations also provide basis for lexical choice. Despite of these advantages, the syntactic information in WordNet is relatively poor. Conversely, the result we obtained after combining COMLEX and EVCA has rich syntactic information, but this information is provided at word level thus unsuitable to use for generation directly. These complementary resources are therefore combined in the second stage, where the subcategorizations and alternations from COMLEX/EVCA for each word are assigned to each sense of the word.

Each synset in WordNet is linked with a list of verb frames, each of which represents a simple syntactic pattern and general semantic constraints on verb arguments, e.g., *Somebody –s something*. The fact that WordNet contains this syntactic information(albeit poor) makes it possible to link the result from COMLEX/EVCA with WordNet.

The merging operation is based on a compatibility matrix, which indicates the compatibility of each subcategorization in COMLEX/EVCA with each verb frame in WordNet. The sub-

[2] The verb can take a prepositional phrase

[3] The verb can take a prepositional phrase, and the subject of the prepositional phrase is the same as the verb's

[4] 2,947 words appear in both resources.

categorizations and alternations listed in COMLEX/EVCA for each word is then assigned to different senses of the word based on their compatibility with the verbs frames listed under that sense of the word in WordNet. For example, if for a certain word, the subcategorizations PP-PRED-RS and NP are listed for the word in COMLEX/EVCA, and the verb frame *somebody –s PP* is listed for the first sense of the word in WordNet, then PP-PRED-RS will be assigned to the first sense of the word while NP will not. We also keep in the lexicon the general constraint on verb arguments from WordNet frames. Therefore, for this example, the entry for the first sense of w indicates that the verb can take a prepositional phrase as a complement, the subject of the verb is the same as the subject of the prepositional phrase, and the subject should be in the semantic category "somebody". As you can see, the result incorporates information from three resources and but is more informative than any of them. An alternation is considered applicable to a word sense if both alternate patterns have matchable verb frames under that sense.

The compatibility matrix is the kernel of the merging operations. The 147*35 matrix (147 subcategorizations from COMLEX/EVCA, 35 verb frames from WordNet) was first manually constructed based on human understanding. In order to achieve high accuracy, the restrictions to decide whether a pair of labels are compatible are very strict when the matrix was first constructed. We then use regressive testing to adjust the matrix based on the analysis of merging results. During regressive testing, we first merge WordNet with COMLEX/EVCA using current version of compatibility matrix, and write all inconsistencies to a log file. In our case, an inconsistency occurs if a subcategorization or alternation in COMLEX/EVCA for a word can not be assigned to any sense of the word, or a verb frame for a word sense does not match any subcategorization for that word. We then analyze the log file and adjust the compatibility matrix accordingly. This process repeated 6 times until when we analyze a fair amount of inconsistencies in the log file, they are no more due to over-restriction of the compatibility matrix.

Inconsistencies between WordNet and COMLEX/EVCA result unmatching subcategorizations or verb frames. On average, 15% of subcategorizations and alternations for a word can not be assigned to any sense of the word, mostly due to the incompleteness of syntactic information in WordNet; 2% verb frames for each sense of a word does not match any subcategorizations for the word, either due to incompleteness of COMLEX/EVCA or erroneous entries in WordNet.

The lexicon at this stage is a rich set of subcategorizations and alternations for each sense of a word, coupled with semantic constraints of verb arguments. For 5,920 words in the result after combining COMLEX and EVCA, 5,676 words also appear in WordNet and each word has 2.5 senses on average. After the merging operation, the average number of subcategorizations is refined from 5.2 per verb in COMLEX/EVCA to 3.1 per sense, and the average number of alternations is refined from 1.0 per verb to 0.2 per sense. Figure 3 shows the result for the verb *appear* after the merging operation.

appear:
sense 1 give an impression
```
((PP-TO-INF-RS :PVAL ("to") :SO ((sb, -)))
(TO-INF-RS :SO ((sb, -)))
(NP-PRED-RS :SO ((sb, -)))
(ADJP-PRED-RS :SO ((sb, -) (sth, -)))))
```
sense 2 become visible
```
((PP-TO-INF-RS :PVAL ("to")
    :SO ((sb, -) (sth, -)))
...
(INTRANS THERE-V-SUBJ
         :ALT there-insertion
         :SO ((sb, -) (sth, -))))
...
```
sense 8 have an outward expression
```
((NP-PRED-RS :SO ((sth, -)))
(ADJP-PRED-RS :SO ((sb, -) (sth, -))))
```

Figure 3: Entry for the verb *appear* after merging WordNet with the result from COMLEX and EVCA.

2.3 Corpus analysis

Finally, we enriched the lexicon with language usage information derived from corpus analysis. The corpus used here is the Brown Corpus. The language usage information in the lexicon include: (1) frequency of each word sense; (2) frequency of subcategorizations for each word sense. A parser is used to recognize the subcategorization of a verb. The corpus analysis information complements the subcategorizations from the static resources by marking potential superfluous entries and supplying entries that are possibly missing in the lexical databases; (3) semantic constraints of verb arguments. The arguments of each verb are clustered based on hyponymy hierarchy in WordNet. The semantic categories we thus obtained are more specific compared to the general constraint(animate or inanimate) encoded in WordNet frame representation. The language usage information is especially useful in lexical choice.

2.4 Discussion

Merging resources is not a new idea and previous work has investigated integration of resources for machine translation and interpretation (Klavans et al., 1991), (Knight and Luk, 1994). Whereas our work differs from previous work in that for the first time, a generation lexicon is built by this technique; unlike other work which aims to combine resources with similar type of information, we select and combine multiple resources containing different types of information; while others combine not well formatted lexicon like LDOCE (Longman Dictionary of Contemporary English), we chose well formatted resources (or manually format the resource) so as to get reliable and usable results; semi-automatic rather than fully automatic approach is adopted to ensure accuracy; corpus analysis based information is also linked with information from static resources. By these measures, we are able to acquire an accurate, reusable, rich, and large-scale lexicon for natural language generation.

3 Applications

3.1 Architecture

We applied the lexicon to lexical choice and lexical realization in a practical generation system. First we introduce the architecture of lexical choice and realization and then describe the overall system.

A multi-level feedback architecture as shown in Figure 4 was used for lexical choice and realization. We distinguish two types of concepts: semantic concepts and lexical concepts. A semantic concept is the semantic meaning that a user wants to convey, while a lexical concept is a lexical meaning that can be represented by a set of synonymous words, such as synsets defined in WordNet. Paraphrases are also distinguished into 3 types according to whether they are at the semantic, lexical, or syntactic level. For example, if asked whether you will be at home tomorrow, then the answers “I’ll be at work tomorrow”, “No, I won’t be at home.’, and “I’m leaving for vacation tonight” are paraphrases at the semantic level. Paraphrases like “He bought an umbrella” and “He purchased an umbrella” are at the lexical level since they are acquired by substituting certain words with synonymous words. Paraphrases like “A ship appeared on the horizon” and “On the horizon appeared a ship” are at the syntactic level since they only involve syntactic transformations. Therefore, all paraphrases introduced by alternations are at syntactic level. Our architecture includes levels corresponding to these 3 levels of paraphrasing.

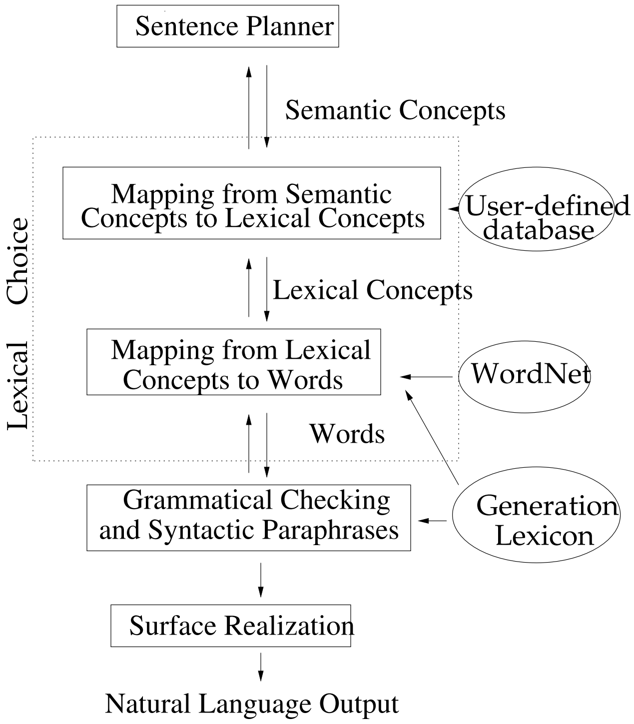

Figure 4: The Architecture for Lexical Choice and Realization

The input to the lexical choice and realization module is represented as semantic concepts. In the first stage, semantic paraphrasing is carried out by mapping semantic concepts to lexical concepts. Generally, semantic level paraphrases are very complex. They depend on the situation, the domain, and the semantic relations involved. Semantic paraphrases are represented declaratively in a database file which can be edited by the users. The file is indexed by semantic concepts and under each entry, a list of lexical concepts that can be used to realize the semantic concept are provided.

In the second stage, we use the lexical resource that we constructed to choose words for the lexical concepts produced by stage 1. The lexicon is indexed by lexical concepts that point to synsets in WordNet. These synsets represent a set of synonymous words and thus, it is at this stage that lexical paraphrasing is handled. In order to choose which word to use for the lexical concept, we use domain-independent constraints that are included in the lexicon as well as domain-specific constraints. Syntactic constraints that come from the detailed subcategorizations linked to each word sense is a domain-independent constraint. Subcategorizations are used to check that the input can be realized by the word. For example, if the input has 3 arguments, then words which take only 2 arguments can not be selected. Semantic constraints on verb argument derived from WordNet and the corpus are used to check the agreement of the arguments. For example, if the input subject argument is an animate, then words which take only inanimate subject can not be selected. Frequency information derived from the corpus is also used to constrain word choice. Besides the above domain-independent constraints other constraints specific to a domain might also be needed to choose an appropriate word for the lexical concept. Introducing the combined lexicon at this stage allows us to produce many lexical paraphrases without much effort; it also allows us to separate domain-independent and domain-specific constraints in lexical choice so that domain-independent constraints can be reused in each application.

The third stage produces a structure represented as a high level sentence structure, with subcategorizations and words associated with each sentence. At this stage, information in the lexical resource about subcategorization and alternations are applied in order to generate syntactic paraphrases. Output of this stage is then fed directly to the surface realization package, the FUF/SURGE system (Elhadad, 1992; Robin, 1994). To choose which alternate pattern of an alternation to use, we use information such as focus of the sentence as criteria; when the two alternates are not distinctively different, such as "He knocked the door" and "He knocked at the door", one of them is randomly chosen. The application of subcategorizations in the lexicon at this stage helps to check that the output is grammatically correct, and alternations can produce many syntactic paraphrases.

The above refining processing is interactive. When a lower level can not find a possible candidate to realize the high level representation, feedback is sent to the higher level module, which then makes changes accordingly.

3.2 PlanDOC

Using the proposed architecture, we applied the lexicon to a practical generation system, PlanDOC. PlanDOC is an enhancement to Bellcore's LEIS-PLAN^{TM} network planning product. It transforms lengthy execution traces of engineer's interaction with LEIX-PLAN into human-readable summaries.

For each message in PlanDOC, at least 3 paraphrases are defined at semantic level. For example, "The base plan called for one fiber activation at CSA 2100" and "There was one fiber activation at CSA 2100" are semantic paraphrases in PlanDOC domain. At the lexical level, we use synonymous words from WordNet to generate lexical paraphrases. A sample lexical paraphrase for "The base plan called for one fiber activation at CSA 2100" is "The base plan proposed one fiber activation at CSA 2100". Subcategorizations and alternations from the lexicon are then applied at the syntactic level. After three levels of paraphrasing, each message in PlanDOC on average has over 10 paraphrases.

For a specific domain such as PlanDOC, an enormous proportion of a general lexicon like the one we constructed is unrelated thus unused at all. On the other hand, domain-specific knowledge may need to be added to the lexicon. The problem of how to adapt a general lexicon to a particular application domain and merge domain ontologies with a general lexicon is out of the scope of this paper but discussed in (Jing, 1998).

4 Conclusion

In this paper, we present research on building a rich, large-scale, and reusable lexicon for generation by combining multiple heterogeneous linguistic resources. Novel semi-automatic transformation and integration were used in combining resources to ensure reliability of the resulting lexicon. The lexicon, together with a multi-level feedback architecture, is used in a practical generation system, PlanDOC.

The application of the lexicon in a generation system such as PlanDOC has many advantages. First, paraphrasing power of the system can be greatly improved due to the introduction of synonyms at the lexical concept level and alternations at the syntactic level. Second, the integration of the lexicon and the flexible architecture enables us to separate the domain-dependent component of the lexical choice module from domain-independent components so they can be reused. Third, the integration of the lexicon with the surface realization system helps in checking for grammatical errors and also simplifies the interface input to the realization system. For these reasons, we were able to develop PlanDOC system in a short time.

Although the lexicon was developed for generation, it can be applied in other applications too. For example, the syntactic-semantic constraints can be used for word sense disambiguation (Jing et al., 1997); The subcategorization and alternations from EVCA/COMLEX are better resources for parsing; WordNet enriched with syntactic information might also be of value to many other applications.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. IRI 96-19124, IRI 96-18797 and by a grant from Columbia University's Strategic Initiative Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Michael Elhadad. 1992. *Using Argumentation to Control Lexical Choice: A Functional Unification-Based Approach.* Ph.D. thesis, Department of Computer Science, Columbia University.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING'94*, Kyoto, Japan.

Hongyan Jing, Vasileios Hatzivassiloglou, Rebecca Passonneau, and Kathleen McKeown. 1997. Investigating complementary methods for verb sense pruning. In *Proceedings of ANLP'97 Lexical Semantics Workshop*, pages 58–65, Washington, D.C., April.

Hongyan Jing. 1998. Applying wordnet to natural language generation. In *To appear in the Proceedings of COLING-ACL'98 workshop on the Usage of WordNet in Natural Language Processing Systems*, University of Montreal, Montreal, Canada, August.

J. Klavans, R. Byrd, N. Wacholder, and M. Chodorow. 1991. Taxonomy and polysemy. Technical Report Research Report RC 16443, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY 10598.

Kevin Knight and Steve K. Luk. 1994. Building a large-scale knowledge base for machine translation. In *Proceedings of AAAI'94.*

H Kučera and W. N. Francis. 1967. *Computational Analysis of Present-day American English.* Brown University Press, Providence, RI.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press, Chicago, Illinois.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, **3**(4):235–312.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. Cognitive Science Laboratory, Princeton University.

Jacques Robin. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation, and Evaluation.* Ph.D. thesis, Department of Computer Science, Columbia University. Also Technical Report CU-CS-034-94.