Stochastic Language Generation in a Dialogue System: Toward a Domain Independent Generator

Nathanael Chambers and **James Allen**
Institute for Human and Machine Cognition
40 South Alcaniz Street
Pensacola, FL 32502
{nchambers,jallen}@ihmc.us

Abstract

Until recently, surface generation in dialogue systems has served the purpose of simply providing a backend to other areas of research. The generation component of such systems usually consists of templates and canned text, providing inflexible, unnatural output. To make matters worse, the resources are typically specific to the domain in question and not portable to new tasks. In contrast, domain-independent generation systems typically require large grammars, full lexicons, complex collocational information, and much more. Furthermore, these frameworks have primarily been applied to text applications and it is not clear that the same systems could perform well in a dialogue application. This paper explores the feasibility of adapting such systems to create a domain-independent generation component useful for dialogue systems. It utilizes the domain independent semantic form of The Rochester Interactive Planning System (TRIPS) with a domain independent stochastic surface generation module. We show that a written text language model can be used to predict dialogue utterances from an over-generated word forest. We also present results from a human oriented evaluation in an emergency planning domain.

1 Introduction

This paper takes steps toward three surface generation goals in dialogue systems; to create a domain-independent surface generator, to create a surface generator that reduces dependence on large and/or domain-specific resources by using out of domain language models, and to create an effective human-like surface generator.

Natural Language systems are relatively young and most of today's architectures are designed and tested on specific domains. It is becoming increasingly desirable to build components that are domain-independent and require a small amount of time to instantiate. Unfortunately, when components are tailored to a specific domain, it requires a complete overhaul to use the architecture in a new domain.

While dialogue systems have found success in many areas, the backend of these systems, Natural Language Generation (NLG), has largely been ignored and used solely to show the progress of other components. However, it is now important to generate not just content-rich utterances, but also *natural* utterances that do not interfere with the dialogue. Easy to build template-based NLG components can usually satisfy the content requirement, but their static, inflexible forms rarely facilitate an effective human oriented dialogue system.

Natural surface generation requires hand-crafted lexicons, grammars, ontologies, and much more to be successful. The time required to create a simple surface generation component is small, but the time required to create even a mildly natural component is very large. Language modeling offers hope that the information encoded in these grammars and lexicons is implicitly present in spoken and written text. There have been many advances with stochastic approaches in areas that have taken advantage of the large corpora of available newswire, such as Machine Translation (MT). If newswire text (which makes up much of the available English corpora) can be applied to dialogue, we could depend less on hand-crafted grammars and domain-specific resources.

This paper describes an approach to surface generation in dialogue systems that uses out of domain language models; a model based on newswire text and a model based on spoken dialogue transcripts. We also describe how this approach fits with a domain independent logical form being used for interpretation in TRIPS. Our analysis of this approach shows that newswire corpora can generate not only the semantic *content* in its output, but

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Stochastic Language Generation in a Dialogue System: Toward a Domain Independent Generator	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Institute for Human and Machine Cognition,40 South Alcaniz Street,Pensacola,FL,32502**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**10**	

also shows that it can be integrated successfully into a dialogue system, resulting in only a slight decrease in *naturalness* as judged by human evaluators.

This paper begins with a description of previous surface generation work. Section 3 describes the stochastic algorithm used from the Machine Translation (MT) system, HALogen, including differences in dialogue versus newswire text. Section 4 describes the domain independence of the logical form in TRIPS and how independence is preserved in translating into the stochastic component. Section 5 describes our evaluation including the language models and the domain we used for evaluation. Finally, we present the results and discussion in section 6.

2 Background

Template-based approaches have been widely used for surface generation. This has traditionally been the case because the many other areas of NLP research (speech recognition, parsing, knowledge representation, etc.) within a dialogue system require an output form to indicate the algorithms are functional. Templates are created very cheaply, but provide a rigid, inflexible output and poor text quality. See Reiter (Reiter, 1995) for a full discussion of templates. Dialogue systems particularly suffer as understanding is very dependent on the naturalness of the output.

Rule-based generation has developed as an alternative to templates. Publicly available packages for this type of generation take strides toward independent generation. However, a significant amount of linguistic information is usually needed in order to generate a modest utterance. This kind of detail is not available to most domain *independent* dialogue systems. A smaller, domain-specific rule-based approach is difficult to port to new domains.

The corpus-based approach to surface generation does not use large linguistic databases but rather depends on language modeling of corpora to predict correct and natural utterances. The approach is attractive in comparison to templates and rule-based approaches because the language models implicitly encode the natural ordering of English. Recently, the results from corpus-based surface generation in dialogue systems have been within specific domains, the vast majority of which have used the Air Travel Domain with Air Travel corpora.

Ratnaparkhi (Ratnaparkhi, 2000; Ratnaparkhi, 2002) and Oh and Rudnicky (Oh and Rudnicky, 2000) both studied surface generators for the air travel domain. Their input semantic form is a set of attribute-value pairs that are specific to the airline reservation task. The language models were standard n-gram approaches that depended on a tagged air travel corpus for the attribute types. Both groups ran human evaluations; Ratnaparkhi studied a 2 subject evaluation (with marks of OK,Good,Bad) and Oh and Rudnicky studied 12 subjects that compared the output between a template generator and the corpus-based approach. The latter showed no significant difference.

Most recently, Chen et al. utilized FERGUS (Bangalore and Rambow, 2000) and attempted to make it more domain independent in (Chen et al., 2002). There are two stochastic processes in FERGUS; a tree chooser that maps an input syntactic tree to a TAG tree, and a trigram language model that chooses the best sentence in the lattice. They found that a domain-specific corpus performs better than a Wall Street Journal (WSJ) corpus for the trigram LM. Work was done to try and use an independent LM, but (Rambow et al., 2001) found interrogatives to be unrepresented by a WSJ model and fell back on air travel models. This problem was not discussed in (Chen et al., 2002). Perhaps automatically extracted trees from the corpora are able to create many good and few bad possibilities that the LM might choose.

(Chen et al., 2002) is the first paper to this author's knowledge that attempts to create a stochastic domain independent generator for dialogue systems. One of the main differences between FERGUS and this paper's approach is that the input to FERGUS is a deep syntactic tree. Our approach integrates semantic input, reducing the need for large linguistic databases and allowing the LM to choose the correct forms. We are also unique in that we are intentionally using two out-of-domain language models. Most of the work on FERGUS and the previous surface generation evaluations in dialogue systems are dependent on English syntax and word choice within the air travel domain. The final generation system cannot be ported to a new domain without further effort. By creating grammar rules that convert a semantic form, some of these restrictions can be removed. The next section describes our stochastic approach and how it was modified from machine translation to spoken dialogue.

3 Stochastic Generation (HALogen)

We used the HALogen framework (Langkilde-Geary, 2002) for our surface generation. HALogen was originally created for a domain within MT and is a sentence planner and a surface realizer. Analysis and MT applications can be found in (Langkilde and Knight, 1998; Knight and Langkilde, 2000).

HALogen accepts a feature-value structure ranging from high-level semantics to shallow syntax. Figure 1 shows a mixture of both as an example. Given this input, generation is a two step process. First, the input form is converted into a word forest (a more efficient representation of a word lattice) as described in (Langkilde-Geary, 2002). Second, the language model chooses the most probable path through the forest as the output sentence.

```
(V68753 / move
        :TENSE past
        :AGENT (V68837 / person
                :QUANT three
                :NUMBER plural
        )
        :THEME (V68846 / ambulance
        )
)
```

Figure 1: HALogen input of the sentence *Three people moved the ambulance.*

The word forest is created by a series of grammar rules that are designed to over-generate for a given representation. As figure 1 shows, there is a lot of syntactic information missing. The rules are not concerned with generating only syntactically correct possibilities, but to generate all possibilities under every input that is not specified (our example does not provide a determiner for *ambulance*, so the grammar would produce the definite and indefinite versions). Once the forest is created, the language model chooses the best path(s) through the forest.

We modified HALogen's grammar to fit the needs of a dialogue system while maintaining the same set of roles and syntactic arguments recognized by the grammar. The TRIPS Logical Form uses many more roles than HALogen recognizes, but we converted them to the smaller set. By using HALogen's set of roles, we can be assured that our grammar is domain independent from TRIPS. We did, however, expand the grammar within its current roles. For instance, we found the *theme* role to be insufficient and changed the grammar to generate more syntactic constructs (for example, we generate the theme in both the object and subject positions). We also expanded the production rules for interrogatives and imperatives, both of which were sparsely used/tested because of HALogen's original use in MT domains.

HALogen is able to expand WordNet word classes into their lexical items, but due to the difficulty of mapping the TRIPS word classes to WordNet, our input terms to HALogen are the desired lexical items instead of word classes as shown in figure 1. Future work includes linking the grammar to the TRIPS word classes instead of WordNet.

4 The Dialogue System

We developed our approach within TRIPS, a collaborative planning assistant that interacts with a human user mainly through natural language dialogue, but also through graphical displays. The system supports many domains involving planning scenarios, such as a 911 disaster rescue assistant and a medical adviser. TRIPS performs advanced reasoning and NLP tasks including, but not limited to, interpretation in context, discovering user intentions, planning, and dialogue management. Language generation has largely been ignored in the system until recently. As with many dialogue systems, it has simply been a means to show results in the above areas through a language back-end. Recently, Stent (Stent, 1999) did extensive work on dialogue management through rule-based generation (Allen et al., 2001).

```
(define-type LF_CONSUME
  :semfeatures
     (Situation (aspect dynamic) (cause agentive))
  :arguments
     (AGENT (Phys-obj (intentional +) (origin living)))
     (THEME (Phys-obj (form substance))))
```

Figure 2: LF type definitions for LF_CONSUME (from (Dzikovska et al., 2003))

4.1 Logical Form of Meaning

There are two meaning representations in TRIPS. The first is a domain independent representation called the *logical form* (LF). The second is a domain dependent *knowledge representation* (KR). The effort toward creating the domain independent LF is part of an overall goal of creating a dialogue system that is easily portable to new domains. A domain-specific representation is always needed for reasoning, and mapping rules are created to map the LF into the KR for each domain. These rules are easier to create than a new logical representation for each domain.

Dzikovska, Swift and Allen (Dzikovska et al., 2003) have built a parser that parses speech utterances into this domain-independent LF. The LF is very important to this paper. One of the biggest problems that any surface generation approach faces is that it takes a lot of work to generate sentences for one domain. Moving to a new domain usually involves duplicating much of this work. However, if we create a surface generator that uses the LF as input, we have created a surface generator that is able to generate English in more than one specific domain.

The LF ontology consists of a single-inheritance hierarchy of frame-like LF types that classify entities according to their semantics and argument structure. Every LF type can have a set of thematic arguments with selectional restrictions. The ontology is explicitly designed to capture the semantic differences that can affect sentence structure, and it draws from many sources including FRAMENET, EURO-WORDNET, and VERBNET. The reader is referred to (Dzikovska et al., 2003) for more details. An example of an LF type definition is shown in Figure 2.

```
(SPEECHACT sa1 SA_TELL :content V11)
(F V11 (* LF_CONSUME take) :AGENT V123
                           :THEME V433)
                           :TMA ((:TENSE PAST))
(PRO V123 (:* LF_PERSON he) :CONTEXT_REL HE)
(A V433 (:* LF_DRUG aspirin))
```

Figure 3: Logical Form for the sentence, *he took an aspirin*.

The parser uses the LF type definitions to build a general semantic representation of the input. This is a flat and unscoped representation of the semantics of the sentence that serves as input to the TRIPS discourse interpretation modules (which perform reference resolution, disambiguation, intention recognition to produce the final intended meaning). Figure 3 gives an example of the LF representation of the sentence, *he took an aspirin*. It can be read as follows: A speech act of type SA_TELL occurred with content being V11, which is a proposition of type LF_CONSUME (more specifically "take"), with AGENT V123 and THEME V433. V123 is pronominal form of type LF_PERSON and pro-type HE, and V433 is an indefinitely specified object that is of type LF_DRUG (more specifically "aspirin").

The LF representation serves as the input to our surface generation grammar after a small conversion. If natural human quality dialogue can be produced from this LF, not only has a domain independent generator been created, but also a generator that shares ontologies and lexicons with the parser.

4.2 Integrating HALogen into TRIPS

The task of converting our independent Logical Form (LF) into HALogen's Abstract Meaning Representation was relatively straightforward. Several rules were created to change LF specific roles into the smaller set of roles that the surface generation grammar recognizes. LF roles such as COGNIZER and ENTITY are converted to AGENT and THEME respectively. Verb properties represented by TMA are converted into the appropriate syntactic roles of TENSE, MODALITY, AUXILLARY, etc. The LF type triple is reduced to just the lexical item and appropriate determiners are attached when the LF provides enough information to warrant it. It is best illustrated by example using our example LF in figure 3. Given these decisions, our example's conversion becomes:

```
(V11 / TAKE
     :TENSE PAST
     :AGENT (V123 / HE)
     :THEME (V433 / ASPIRIN))
```

This resulting AMR is the input to HALogen where it is converted into a word forest using our modified dialogue-based HALogen grammar. Finally, the language model chooses the best output.

The above conversion applies to declarative, imperative and interrogative speech acts. These are translated and generated by the method in section 3. We also take a similar approach to Stent's previous work (Stent, 1999) that generated grounding and turn-taking acts using a template-based method. These usually short utterances do not require complex surface generation and are left to templates for proper production.

5 Evaluation

This paper is evaluating two surface generation design decisions: the effectiveness of stochastic (word forest based) surface generation with domain independent language models, and the benefits of using dialogue vs. newswire models. Evaluating any natural language generation system involves many factors, but we focused on two of the most important aspects to evaluate, the content and clarity (naturalness) of the output (English utterances). This section briefly describes previous automatic evaluation approaches that we are avoiding, followed by the human evaluation we have performed on our system.

5.1 Automatic Evaluation

Evaluating generation is particularly difficult due to the diverse amount of *correct* output that can be generated. There are many ways to present a given semantic representation in English and what determines quality of content and form are often subjective measures. There are two general approaches to a surface generation evaluation. The first uses human evaluators to score the output with some pre-defined ranking measure. The second uses a quantitative automatic approach usually based on n-gram presence and word ordering. Bangalore et al. describe some of the quantitative measures that have been used in (Bangalore et al., 2000). Callaway recently used quantitative measures in an evaluation between symbolic and stochastic surface generators in (Callaway, 2003).

The most common quantitative measure is Simple String Accuracy. This metric uses an ideal output string and compares it to a generated string using a metric that combines three word error counts; insertion, deletion, and substitution. One variation on this approach is tree-based metrics. These attempt to better represent how *bad* a bad result is. The tree-based accuracy metrics do not compare two strings directly, but instead build a dependency tree for the ideal string and attempt to create the same dependency tree from the generated string. The score is dependent not only on word choice, but on positioning at the phrasal level. Finally, the most recent evaluation metric is the Bleu Metric from IBM(Papineni et al., 2001).

Designed for Machine Translation, it scores generated sentences based on the n-gram appearance from multiple ideal sentences. This approach provides more than one possible realization of an LF and compares the generated sentence to all possibilities.

Unfortunately, the above automatic metrics are very limited in mimicking human scores. The Bleu metric can give reasonable scores, but the results are not as good when only one human translation is available. These automatic metrics all compare the desired output with the actual output. We decided to ignore this evaluation because it is too dependent on syntactic likeness. The following two sentences represent the same semantic meaning yet appear very different in structure:

The injured person is still waiting at the hospital.
The person with the injury at the hospital is still waiting.

The scoring metrics would judge very harshly, yet a human evaluator should see little difference in semantic content. Clearly, the first is indeed better in naturalness (closeness to human English dialogue), but both content and naturalness cannot be measured with the current quantitative (and many human study) approaches. Although it is very time consuming, human evaluation continues to be the gold standard for generation evaluation.

5.2 Evaluation Methodology

Our evaluation does not compare an *ideal* utterance with a generated one. We use a real human-human dialogue transcript and replace every utterance of *one* of the participants with our generated output. The evaluators are thereby reading a dialogue between a human and a computer generated human, yet it is based on the original human-human dialogue. Through this approach, we can present the evaluators with both our generated and the original transcripts (as the control group). However, they do not know which is artificial, or even that any of them are not human to human. The results will give an accurate portrayal of how well the system generates dialogue. The two aspects of dialogue that the evaluators were asked to measure for each utterance were **understandability** (semantically within context) and **naturalness**.

There have been many metrics used in the past. Metrics range from scoring each utterance with a subjective score (Good,Bad) to using a numeric scale. Our evaluators use a numeric scale from 0 to 5. The main motivation for this is so we can establish averages and performance results more easily. The final step is to obtain a suitable domain of study outside the typical air travel domain.

5.3 Domain Description and Dialogue Construction

A good dialogue evaluation is one in which all aspects of a natural dialogue are present and the only aspect that has been changed is how the surface generation presents the required information. By replacing one speaker's utterances with our generated utterances in a transcript of a real conversation, we guarantee that grounding and turn-taking are still present and our evaluation is not hindered by poor dialogue cues. The TRIPS Monroe Corpus (Stent, 2000) works well for this task.

There are 20 dialogues in the Monroe Corpus. Each dialogue is a conversation between two English speakers. Twenty different speakers were used to construct the dialogues. Each participant was given a map of Monroe County, NY and a description of a task that needed to be solved. There were eight different disaster scenarios ranging from a bomb attack to a broken leg and the participants were to act as emergency dispatchers (this domain is often referred to as the 911 Rescue Domain). One participant U was given control of solving the task, and the other participant S was told that U had control. S was to assist U in solving the task. At the end of the discussion, U was to summarize the final plan they had created together.

The average dialogue contains approximately 500 utterances. We chose three of the twenty dialogues for our evaluation. The three were the shorter dialogues in length (Three of the only four dialogues that are less than 250 utterances long. Many are over 800 utterances.). This was needed for practical reasons so the evaluators could conduct their rankings in a reasonable amount of time and still give accurate rankings. The U and S speakers for each dialogue were different.

We replaced the S speaker in each of the dialogues with generated text, created by the following steps:

- Parse each S utterance into its LF with the TRIPS parser.
- Convert the LF to the AMR grammar format.
- Send the AMR to HALogen.
- Generate the top sentence from this conversion using our chosen LM.

We hand-checked for correctness each AMR that is created from the LF. The volatile nature of a dialogue system under development assured us that many of the utterances were not properly parsed. Any errors in the AMR were fixed by hand and hand constructed when no parse could be made. The fixes were done before we tried to generate the S speaker in the evaluation dialogues.

We are assuming perfect input to generation. This evaluation does not evaluate how well the conversion from

the LF to the AMR is performing. Our goal of generating natural dialogue from a domain-independent LM can be fully determined by analyzing the stochastic approach in isolation. Indeed, the goal of a domain independent generator is somewhat dependent on the conversion from our domain independent LF, but we found that the errors from the conversion are not methodological errors. The errors are simple lexicon and code errors that do not relate to domain-specifics. Work is currently underway to repair such inconsistencies.

Each of the S participant's non-dialogue-management utterances were replaced with our generated utterances. The grounding, turn-taking and acknowledgment utterances were kept in their original form. We plan on generating these latter speech acts with templates and are only testing the stochastic generation in this evaluation. The U speaker remained in its original state. The control groups will identify any bias that U may have over S (i.e. if U speaks 'better' than S in general), but testing the generation with the same speaker allows us to directly compare our language models.

5.4 Language Model Construction

We evaluated two language models. The first is a news source model trained on 250 million words with a vocabulary of 65,529 from the WSJ, AP and other online news sources as built in (Langkilde-Geary, 2002). This model will be referred to as the WSJ LM. The second language model was built from the Switchboard Corpus (J. Godfrey, 1992), a corpus of transcribed conversations and not newswire text. The corpus is comprised of 'spontaneous' conversations recorded over the phone, including approximately 2 million words with a vocabulary of 20,363. This model will be referred to as the SB LM. Both models are trigram, open vocabulary models with Witten-Bell smoothing. The Switchboard Corpus was used because it contrasts the newswire corpus in that it is in the genre of dialogue yet does not include the Monroe Corpus that the evaluation was conducted on.

5.5 Evaluators

Ten evaluators were chosen, all were college undergraduates between the ages of 18-21. None were linguistics or computer science majors. Each evaluator received three transcripts, one from each of our three chosen dialogues. One of these three was the original human to human dialogue. The other two had the S speaker replaced by our surface generator. Half of the evaluators received generations using the WSJ LM and the other half received the SB LM. They ranked each utterance for understandability and naturalness on scales between 0 and 5. A comparison of the human and generated utterances is given in figure 8 in the appendix.

Percent Difference between U and S speakers

	0	1	2	3	4
understand	0.92	6.03	3.70	0.23	1.74
natural	-1.31	-0.26	2.56	1.94	-3.09
	5	6	7	8	9
understand	3.91	3.27	2.46	-0.10	14.8
natural	3.60	2.38	-0.26	5.16	13.3

Total Percent Difference

understand	3.24%
natural	1.85%

Figure 4: Difference between the human evaluator scores for the two original human speakers, U and S. The ten evaluators are listed by number, 0 to 9. Evaluators rated the content (understandability) and clarity (naturalness) of each utterance on a 0-5 scale. S was rated slightly higher than U.

6 Results

Figure 4 compares the control dialogues as judged by the human evaluators by giving the percent difference between the two human speakers. It is apparent that the U speaker is judged worse than the S speaker in the average of the three dialogues. We see the S speaker is scored 3.24% higher in understanding and 1.85% higher in naturalness. Due to the nature of the domain, the U speaker tends to make more requests and short decisions while the S speaker gives much longer descriptions and reasons for his/her actions. It is believed the human evaluators tend to score shorter utterances more harshly because they aren't 'complete sentences' as most people are used to seeing in written text. We believe this also explains the discrepancy of evaluator 9's very high scores for the S speaker. Evaluator 9 received dialogue 10 as his control dialogue. Dialogue 10's S speaker tended to have much longer utterances than any of the other five speakers in the three dialogues. It is possible that this evaluator judged shorter utterances more harshly.

Figure 5 shows the comparison between using the two LMs as well as the human control group. The scores shown are the average utterance scores over all evaluators and dialogues. The dialogue management (grounding, turn-taking, etc.) utterance scores are not included in these averages. Since we do not generate these types of utterances, it would be misleading to include them in our evaluation. As figure 5 shows, the difference between the two LMs is small. Both received a lower naturalness score than understandability. It is clear that we are able to generate utterances that are understood, but yet are slightly less natural than a human speaker.

Figure 6 shows the distribution of speech acts in each of the 3 evaluation dialogues. Due to the nature of the Monroe Corpus, there are not many interrogatives or im-

Language Model Comparison

	U	S	U/S difference
WSJ LM			
understand	4.67	4.33	-0.34 (-7.28%)
natural	4.49	3.97	-0.52 (-11.58%)
SB LM			
understand	4.62	4.30	-0.32 (-6.93%)
natural	4.18	3.84	-0.34 (-8.13%)
HUMAN			
understand	4.63	4.78	0.15 (3.24%)
natural	4.33	4.41	0.08 (1.85%)

Figure 5: Average scores (over the 10 evaluators) of understandability and naturalness with the dialogue management utterances removed. The first compares the S speaker generated with the WSJ LM, the second compares the S speaker generated with the SB LM, and the third is the S speaker using the original human utterances.

peratives. Since the two participants in the dialogues work together and neither has more information about the rescue problem than the other, there are not many questions. Rather, it is mostly declaratives and acknowledgments.

Figure 7 shows the average score given for each speech act across all evaluators. Note that the numbers are only for the S speaker in each dialogue because only S was generated with the surface generator. Since each evaluator scored 2 computer dialogues and 1 human (control) dialogue, the LM numbers are averaged across twice as many examples. The understandability scores for the WSJ and SB LMs are relatively the same across all acts, but naturalness is slightly less in the SB LM. Comparing the human scores to both out-of-domain LMs, we see that declaratives averaged almost a 0.5 point loss from the human control group in both understandability and naturalness. Imperatives suffer an even larger decrease with an approximate 0.7 loss in understandability. The SB LM actually averaged over 1.0 decrease in naturalness. The interrogatives ranged from a 0.5 to 0 loss.

6.1 Discussion

We can conclude from figure 5 that the evaluators were relatively consistent among each other in rating understandability, but not as much so with naturalness. The comparison between the WSJ and SB LMs is inconclusive because we see in figure 5 that even though the evaluators gave the WSJ utterances *higher* absolute scores than the SB utterances, the percent difference from how they ranked the human U speaker is *lower*. The fact that it is inconclusive is somewhat surprising because intuition leads us to believe that the dialogue-based SB would perform better than the newswire-based WSJ. One reason may be because the nature of the Monroe Corpus does not include many dialogue specific acts such as questions and imperatives. However, declaratives are well represented and we can conclude that the newswire WSJ LM is as effective as the dialogue SB model for generating dialogue declaratives. Also, it is of note that the WSJ LM out-performed the SB LM in naturalness for most speech act types (as seen in figure 7) as well.

The main result from this work is that an out-of-domain language model cannot only be used in a stochastic dialogue generation system, but the large amount of available *newswire* can also be effectively utilized. We found only a 7.28% decrease in understandability and an 11.58% decrease in naturalness using our *newswire* LM. This result is exciting. These percentages correspond to ranking an utterance 4.64 and 4.42 instead of a perfect 5.00 and 5.00. The reader is encouraged to look at the output of the generation in the appendix, figure 8.

6.2 Future Work

We have created a new grammar to generate from the LF that recognizes the full set of thematic roles. In addition, we have linked our dialogue system's lexicon to the generation module instead of WordNet, resulting in a fully integrated component to be ported to new domains with little effort. It remains to run an evaluation of this design.

Also, stochastic generation favors other avenues of generation research, such as user adaptation. Work is being done to adapt to the specific vocabulary of the human user using dynamic language models. We hope to create an adaptive, natural generation component from this effort.

Finally, we are looking into random weighting approaches for the generation grammar rules and resulting word forest in order to create dynamic surface generation. One of the problems of template-based approaches is that the generation is too static. Our corpus-based approach solves much of the problem, but there is still a degree of 'sameness' that is generated among the utterances.

7 Conclusion

We have shown that steps toward a domain-independent NLG component of a dialogue system can be taken through a corpus-based approach. By depending on a domain-independent semantic input in combination with a grammar that over-generates possible English utterances and a newswire language model to choose the best, we have shown that it is possible to generate *content rich* and *natural* utterances. We report results in a new, richer domain for stochastic generation research and show our approach resulting in only an 11.6% decrease in naturalness when compared to a human speaker.

	Dialogue Mgmt.	Declarative	Imperative	YN-Question	WH-Question
Dialogue 1	45	75	10	7	3
Dialogue 2	49	84	4	17	8
Dialogue 3	57	81	7	1	1

Figure 6: The number of types of speech acts in each of the three dialogues.

	Dialogue Mgmt.	Declarative	Imperative	YN-Question	WH-Question
WSJ LM					
und	4.92	4.34	3.83	4.39	4.78
nat	4.87	3.96	3.73	3.82	4.11
SB LM					
und	4.63	4.33	4.03	4.31	4.89
nat	4.59	3.87	3.21	4.00	3.33
HUMAN					
und	4.73	4.79	4.71	4.76	4.83
nat	4.74	4.41	4.32	4.51	4.83

Figure 7: Comparison of speech act scores of the S speaker. The numbers are averages over the evaluators' scores on a 0-5 scale.

8 Acknowledgments

We give thanks to Lucian Galescu, Irene Langkilde-Geary and Amanda Stent for helpful comments and suggestions on previous drafts of this paper. This work was supported in part by ONR grant 5-23236.

References

J. Allen, G. Ferguson, and A. Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of Intelligent User Interfaces 2001 (IUI-01)*, Santa Fe, NM, January.

S. Bangalore and O. Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrucken, Germany.

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG 2000)*, Mitzpe Ramon, Israel.

C. Callaway. 2003. Evaluating coverage for large symbolic nlg grammars. In *IJCAI*, pages 811–817, Acapulco, Mexico.

J. Chen, S. Bangalore, O. Rambow, and M. Walker. 2002. Towards automatic generation of natural language generation systems. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

M. Dzikovska, M. Swift, and J. Allen. 2003. Constructing custom semantic representations from a generic lexicon. In *5th International Workshop on Computational Semantics*.

J. McDaniel J. Godfrey, E. Holliman. 1992. Switchboard: Telephone speech corpus for research and development. In *ICASSP*, pages 517–520, San Francisco, CA.

K. Knight and I. Langkilde. 2000. Preserving ambiguities in generation via automata intersection. In *American Association for Artificial Intelligence conference (AAAI)*.

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the ACL/COLING-98*, Montreal, Quebec.

I. Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *International Natural Language Generation Conference (INLG)*.

A. Oh and A. Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In *ANLP/NAACL 2000 Workshop on Conversational Systems*, pages 27–32, May.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Research Report RC22176, IBM, September.

O. Rambow, S. Bangalore, , and M. Walker. 2001. Natural language generation in dialog systems. In *First International Conference on Human Language Technology Research (HLT2001)*.

A. Ratnaparkhi. 2000. Trainable methods for surface natural language generation. In *1st Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2000)*.

A. Ratnaparkhi. 2002. Trainable approaches to surface natural language generation and their application to

conversational dialog systems. *Computer, Speech, & Language*.

E. Reiter. 1995. Nlg vs. templates. In *Fifth European Workshop on Natural-Language Generation*.

A. Stent. 1999. Content planning and generation in continuous-speech spoken dialog systems. In *KI'99 workshop*.

A. Stent. 2000. The monroe corpus. Research Report 728, Computer Science Dept., University of Rochester, March. 99-2.

9 Appendix

__	__	7	right so we have bunch of roads that are out
__	__	*7	we have a bunch of roads that are out
__	__	8	a bunch of electric lines that are down
__	__	*8	bunch of electric lines that are down
__	__	9	and we need to fix them
__	__	*9	and we need to fix them
__	__	10	lets see
__	__	*10	let us see
__	__	11	and one of the electric lines is across the road and we need to fix that
__	__	*11	one electric lines are across the road and we need to fix that immediately
__	__	13	it is across
__	__	*13	it is across
__	__	14	its at the intersection of three eighty three and two fifty two a just
__	__	*14	it was at the intersection of three eighty three and two fifty two as
__	__	16	so
__	__	*16	so
__	__	18	yeah so i want so we need to send an electric crew
__	__	*18	yeah so we need to send electric crews
__	__	19	i guess theres only one set of electric crews
__	__	*19	i guess there is one set of electric crews
__	__	20	uh send them there to shut off the power
__	__	*20	send them the power to shut off in there
__	__	22	and that should take about twenty minutes
__	__	*22	twenty minutes minutes and that should take
__	__	23	um not going to worry about travel time perhaps
__	__	*23	perhaps we will not travel time worry
__	__	24	and then after that i would send the airport road crew to the same location
__	__	*24	i would send the airport crew fixed the road to the same location
__	__	28	i guess
__	__	*28	i guess
__	__	29	but they can shut off the power from an intersection
__	__	*29	they can shut the power of an intersection off
__	__	31	um before that
__	__	*31	before that
__	__	32	okay so thats one location
__	__	*32	okay so that is one location
__	__	33	and its going to take them four hours to fix the road
__	__	*33	and they will take four hours to fix the roads
__	__	35	and then after that we can send an electric crew to um restore the lines
__	__	*35	and then we can send electric crews to restore the lines
__	__	36	which takes two hours
__	__	*36	that takes two hours
__	__	38	six plus twenty minutes yeah
__	__	*38	six minutes plus twenty minutes

Figure 8: A comparison of the original human and our generated utterances, part of dialogue three in the Monroe Corpus (just the S speaker). The starred numbers are inserted into the dialogue to provide a side by side comparison of the quality of our generation. Starred utterances are generated by the approach described in this paper. The evaluators did *not* receive such a dialogue. All human or all generation was presented as the S speaker to each evaluator.