An Unsupervised Method for Learning Generation Dictionaries for Spoken Dialogue Systems by Mining User Reviews

RYUICHIRO HIGASHINAKA
NTT Corporation
MARILYN A. WALKER
University of Sheffield
and
RASHMI PRASAD
University of Pennsylvania

Spoken language generation for dialogue systems requires a dictionary of mappings between the semantic representations of concepts that the system wants to express and the realizations of those concepts. Dictionary creation is a costly process; it is currently done by hand for each dialogue domain. We propose a novel unsupervised method for learning such mappings from user reviews in the target domain and test it in the restaurant and hotel domains. Experimental results show that the acquired mappings achieve high consistency between the semantic representation and the realization and that the naturalness of the realization is significantly higher than the baseline.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language generation*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Natural language interfaces*; H.5.2 [**Information Interfaces and Presentation**]: User interfaces—*Natural language*

General Terms: Languages, Human Factors

Additional Key Words and Phrases: Natural language generation, generation dictionary, user reviews, spoken dialogue systems

This paper is a modified and augmented version of our earlier reports Higashinaka et al. [2005, 2006].
This work was supported by a Royal Society Wolfson Award to M. Walker and a research collaboration grant from NTT to the Cognitive Systems Group at the University of Sheffield.
Authors' addresses: R. Higashinaka, NTT Communication Science Laboratories, NTT Corp., 2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan; email: rh@cslab.kecl.ntt.co.jp; M. A. Walker, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, U.K; email: walker@dcs.shef.ac.uk; R. Prasad, Institute for Research in Cognitive Science, University of Pennsylvania, 3401 Walnut Street, Suite 400A Philadelphia, PA 19104-6228, USA; email: rjprasad@linc.cis.upenn.edu.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2007 ACM 1550-4875/2007/10-ART8 $5.00 DOI 10.1145/1289600.1289601 http://doi.acm.org/10.1145/1289600.1289601

ACM Reference Format:
Higashinaka, R., Walker, M. A., and Prasad, R. 2007. An unsupervised method for learning generation dictionaries for spoken dialogue systems by mining user reviews. ACM Trans. Speech Lang. Process. 4, 4, Article 8 (October 2007), 25 pages. DOI = 10.1145/1289600.1289601 http://doi.acm.org/10.1145/1289600.1289601

1. INTRODUCTION

Spoken dialogue systems are beginning to achieve some commercial success [Heisterkamp 2001; Pieraccini and Lubensky 2005; Feng et al. 2005], but a remaining obstacle to their widespread deployment is the cost involved in handcrafting the spoken language generation module. Spoken language generation requires a dictionary of mappings between the semantic representations of concepts that the system wants to express and realizations of those concepts [Reiter and Dale 2000; Stent et al. 2004]. Dictionary creation is a costly process: an automatic method for creating them would make dialogue technology more scalable.

Generation dictionaries can be based on generation templates or syntactic trees and rules. Generation templates are widely used in many practical systems for simplicity [Seneff and Polifroni 2000; Theune 2003; Higashinaka et al. 2006]. They are composed of pairs of communicative goals/dialogue acts and templates (surface strings with variables). For example, a dialogue act *refer-info-place*, which confirms an information type and a place name in the weather information service domain, can be mapped to the template "Are you interested in the [info=X] in [place=Y]?", where X and Y are provided by the dialogue manager [Higashinaka et al. 2006].

For systems to assert more complex propositions, templates can be extended to incorporate syntactic structures, such as the Deep Syntactic Structure (DSyntS) in Figure 1 [Melčuk 1988], so that aggregation and other syntactic transformations of utterances as well as context-specific prosody assignment can be realized [Stent et al. 2004; Moore et al. 2004]. Generation rules are also used in systems where semantic representations have complex or hierarchical structures bearing various arguments. The generation process works by converting input such as E-forms [Goddeau et al. 1996] or frames [Bobrow et al. 1977] through the application of rewrite rules in a cascading manner [Baptist and Seneff 2000].

Whether templates or rules are employed, it is widely acknowledged that creating and maintaining good quality mappings is a costly process. This is because dialogue systems, especially those that are task-oriented, have to convey the information requested by users as accurately as possible and in a natural enough form for users to be able to intuitively converse with such systems to access information. To satisfy these objectives, both significant human effort in corpus study and linguistic expertise have been necessary [Reiter and Dale 2000; Reiter et al. 2003; Reiter and Sripada 2002].

Recently, corpus-based approaches have been proposed as one way to reduce the effort involved in developing natural language generators (NLGs). This work is based on over-generating many candidate utterances and then ranking them, using either user feedback or corpus models to generate automatic

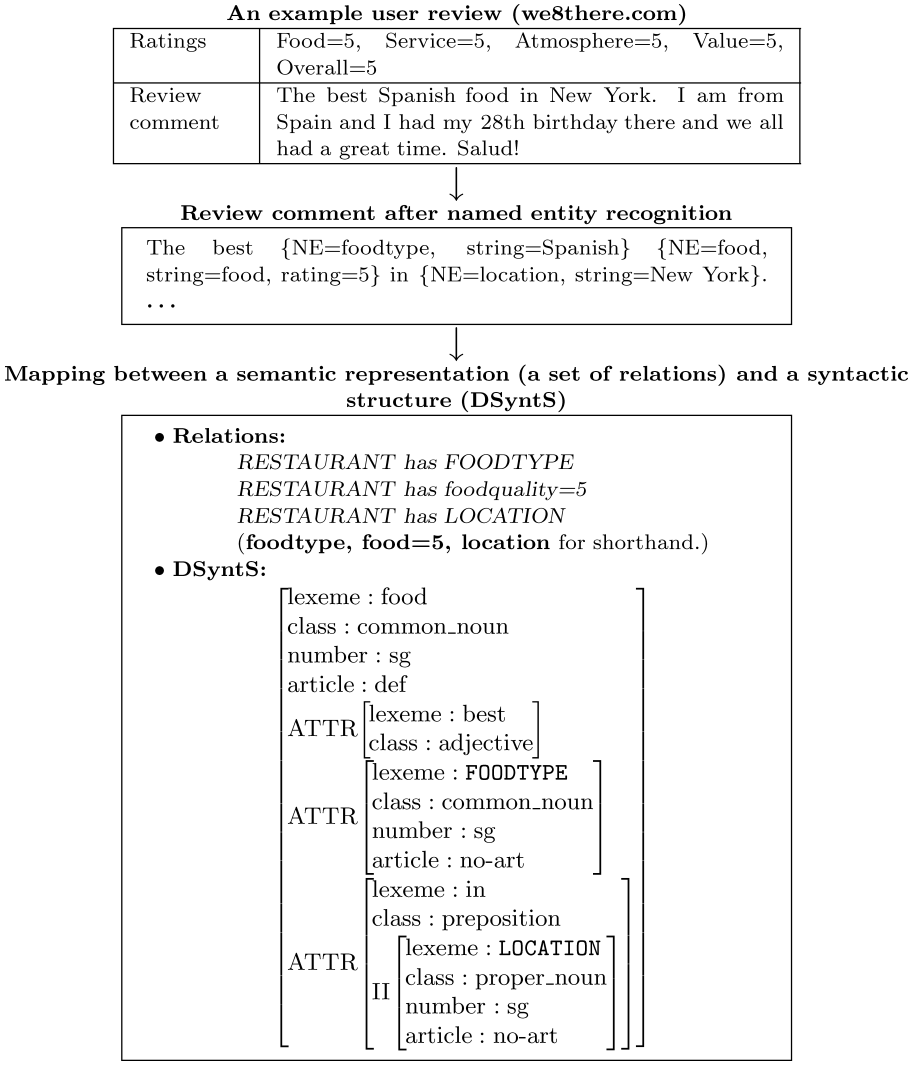

Fig. 1. Example of the procedure for acquiring a generation dictionary mapping.

ranking functions [Langkilde-Geary 2002; Oh and Rudnicky 2000; Rambow et al. 2001; Walker et al. 2002; Stent et al. 2004; Chen et al. 2002; Chambers and Allen 2004]. The advantage of this approach is that the candidate utterances can be generated through many combinations of potential generation procedures without needing to guarantee that every candidate is high quality. However, candidate utterances are still generated with a handcrafted dictionary.

To date, little work has been done on automatically acquiring generation dictionaries which requires that an accurate semantic representation be derived for sentences in the target domain. The related problem of automatically finding sentences with the same meaning has been extensively studied in the field of

automatic paraphrasing. This work typically uses parallel corpora and corpora with multiple descriptions of the same events to extract interchangeable sentences [Barzilay and McKeown 2001; Barzilay and Lee 2003]. Other work has found predicates of similar meanings by using the similarity of contexts around the predicates [Lin and Pantel 2001]. These studies find a set of sentences with the same meaning but do not associate a specific meaning with the sentences.

An exception is work by Barzilay and Lee [2002], Barzilay and Lapata [2006, 2005], and Snyder and Barzilay [2007] which is more in the spirit of our own work. This work is based on supervised techniques with parallel corpora consisting of both complex semantic input and corresponding natural language verbalizations in the domains of mathematical proofs or sports writing. It derives mappings between semantic representations and realizations and explores the training of automatic content selection and aggregation algorithms using these mappings. However, since this technique requires parallel corpora or previously existing semantic transcriptions or labeling, it does not address the problem of the cost involved with dictionary creation.

Other research has begun to explore automatically obtaining semantic representations corresponding to particular linguistic phrases, but this research has not considered whether the learned phrases can be used to generate coherent utterances from the semantic representations, and it has been primarily focused on a small set of semantic relations, such as *is-a* or *part-of* [Pantel and Ravichandran 2004; Gildea and Jurafsky 2002; Etzioni et al. 2005], although recent work has begun to expand beyond this small set of relations [Soderland 2007].

However, when we look at the increasing number of language resources available on the Web, we notice that some of them have specific structures that might be used to facilitate the automatic understanding of the content. For example, tables and lists marked by words such as "pros" and "cons" can be used to collect positive/negative sentence instances for training sentiment classifiers [Kaji and Kitsuregawa 2006], and good/bad votes such as those at amazon.com have been found useful for determining the characteristics of helpful texts [Kim et al. 2006]. Previous work on mining user reviews aim at summarizing reviews so that users can make decisions easily. This work can find adjectives to describe products [Hu and Liu 2005], and automatically find features of a product together with the polarity of adjectives used to describe them [Popescu and Etzioni 2005]. Thus we hypothesize that it may be possible to use the structured information typically available in user review Web sites to induce semantic representations for review sentences.

In this article, we propose a novel method for mining user reviews on the Web to automatically acquire a domain-specific generation dictionary for information presentation in a spoken dialogue system and apply it to user reviews in the restaurant and hotel domains. Our hypothesis is that reviews that provide individual ratings for various distinguished attributes of review entities can be used to map review sentences to semantic representations. Figure 1 shows a user review in the restaurant domain where we hypothesize that the user rating *food* = *5* indicates that the semantic representation for the sentence "The best Spanish food in New York" includes the relation 'RESTAURANT *has foodquality* = *5*.'

In Section 2, we describe our method in detail. Section 3 describes an evaluation experiment based on the application of the method to the two domains. We present results of both objective and subjective evaluations using the learned dictionaries to generate recommendations for hotels and restaurants. Section 4 concludes and describes future work.

2. METHOD: LEARNING A GENERATION DICTIONARY

We propose mining user reviews on the Web to automatically acquire a domain-specific generation dictionary for information presentation in a spoken dialogue system. The basic idea is that the ratings given to the reviews indicate the meaning of sentences, making it possible to derive accurate semantic representations for sentences that are then used to automatically create or augment a generation dictionary.

The automatically created generation dictionary consists of triples $(\mathcal{U}, \mathcal{R}, \mathcal{S})$ representing a mapping between the original utterance $\mathcal{U}$ in the user review, its semantic representation $\mathcal{R}(\mathcal{U})$, and its syntactic structure $\mathcal{S}(\mathcal{U})$. Syntactic structures are derived so that the dictionary can be used in grammar-based full-NLG systems [Stent et al. 2004]. The procedure is outlined briefly in Figure 1. It is comprised of the following steps.

(1) Collect user reviews on the Web to create a population of utterances $\mathcal{U}$.
(2) To derive semantic representations $\mathcal{R}(\mathcal{U})$,
    —identify distinguished attributes and construct a domain ontology;
    —specify lexicalizations of attributes;
    —scrape Web pages' structured data for named entities;
    —tag named entities.
(3) Derive syntactic representations $\mathcal{S}(\mathcal{U})$.
(4) Filter inappropriate mappings.
(5) Add mappings $(\mathcal{U}, \mathcal{R}, \mathcal{S})$ to dictionary.

In what follows, we describe each step in detail.

2.1 Collecting User Reviews

Although there are many Web sites dealing with user reviews, we select only those that have individual ratings for various distinguished attributes of review entities. We collect user reviews from such Web sites and store them as a corpus $(\mathcal{U})$. Some Web sites may contain tabular data for review entities such as names and addresses of restaurants. In such cases, we also store them as additional data.

2.2 Deriving Semantic Representations

We first identify distinguished attributes for each review entity. They include attributes that the users are asked to rate, which have scalar values, and other attributes that can be extracted from the tabular data, which have categorical values. For example, in the restaurant domain, *food, service, atmosphere, value*, and *overall* are scalar-valued distinguished attributes, and *foodtype* and

location are the distinguished attributes with categorical values. Given the distinguished attributes, a simple domain ontology can be automatically derived by assuming that a meronymy relation, represented by the predicate "has" holds between the entity type (e.g., RESTAURANT) and the distinguished attributes. Thus, in the restaurant domain, the following ontology consisting of the seven relations can be derived.

$$\begin{cases} \text{RESTAURANT has foodquality} \\ \text{RESTAURANT has servicequality} \\ \text{RESTAURANT has valuequality} \\ \text{RESTAURANT has atmospherequality} \\ \text{RESTAURANT has overallquality} \\ \text{RESTAURANT has foodtype} \\ \text{RESTAURANT has location} \end{cases}$$

We assume that, although users may discuss other attributes of the entity, at least some of the utterances in the reviews realize the relations specified in the ontology. Our task then is to identify these utterances. We test the hypothesis that if an utterance $\mathcal{U}$ contains named entities corresponding to the distinguished attributes, $\mathcal{R}$ for that utterance includes the relation concerning that attribute in the domain ontology.

We also hypothesize that the rating given for the distinguished attribute specifies the scalar value of the relation. For example, a sentence containing named entities for *foodquality* (e.g., *food* or *meal*. See Table I.) is assumed to realize the relation "RESTAURANT *has foodquality*", and the value of the *foodquality* attribute is assumed to be the value specified in the user rating for that attribute, for example, "RESTAURANT *has foodquality* = 5" in Figure 1. Similarly, the other relations in Figure 1 are assumed to be realized by the utterance "The best Spanish food in New York" because it contains one FOODTYPE named entity and one LOCATION named entity. Values of categorical attributes are replaced by variables representing their type before the learned mappings are added to the dictionary, as shown in Figure 1.

To detect our distinguished attributes, we prepare a named-entity tagger. Currently, lexicalizations of rating-related distinguished attributes are created by hand,[1] and those for other distinguished attributes are automatically imported from the tabular data. We also prepare lexicalizations of named entities that are not relevant to the domain from Web pages in order to detect pieces of information irrelevant to the domain. Finally, we augment the named-entity tagger with our list of lexicalizations, and apply it to the review sentences to derive semantic representations.

2.3 Parsing and DSyntS Conversion

We adopt DSyntSs as a format for syntactic structures because they can be realized by the fast portable realizer RealPro [Lavoie and Rambow 1997]. Since DSyntSs are a type of dependency structure, we first process the sentences with Minipar [Lin 1998], a general-purpose dependency parser, and then convert

[1]In the future, we will investigate other techniques for bootstrapping the lexicalizations.

Minipar's representations into DSyntSs with a converter we developed. We also apply a POS tagger [Brill 1992] to the sentences in parallel so that the converter can incorporate the Penn Tree POS tag information which is not in the Minipar output.

The conversion process consists of three parts. The first part reads in the dependencies shown in the Minipar output as a tree. Two hand-constructed mapping tables are used. One maps the Minipar dependency arc labels to those used by the RealPro DSyntS, and the other maps the POS tags to feature structures for the RealPro DSyntS. This results in the assignment of dependency labels and feature structures to each node in the tree. The second part uses handcrafted rules to modify the derived tree structures for easily identifiable inaccuracies in dependencies and feature structures. Finally, the tree is converted to the representation that is used by the RealPro surface realizer [Lavoie and Rambow 1997].

Since we are processing sentences from user reviews, which are different from the newspaper articles on which Minipar was trained, the output of Minipar for such sentences can be inaccurate, leading to failure in conversion. We check whether the conversion is successful for each sentence in the filtering stage.

2.4 Filtering

The goal of filtering is to identify $\mathcal{U}$ that realizes the distinguished attributes and to guarantee high precision for the learned mappings. Recall is less important since systems need to convey requested information as accurately as possible. Our procedure for deriving semantic representations is based on the hypothesis that if $\mathcal{U}$ contains named entities that realize the distinguished attributes, $\mathcal{R}$ will include the relevant relation in the domain ontology.

We also assume that, if $\mathcal{U}$ contains named entities that are not covered by the domain ontology or contains words indicating that the meaning of $\mathcal{U}$ depends on the surrounding context, $\mathcal{R}$ will not completely characterize the meaning of $\mathcal{U}$, and so $\mathcal{U}$ should be eliminated. We also require an accurate $\mathcal{S}$ for $\mathcal{U}$. Therefore, the filters described in the following eliminate $\mathcal{U}$ that (1) realizes semantic relations not in the ontology; (2) contains words indicating that its meaning depends on the context; (3) contains unknown words; or (4) cannot be parsed accurately. The filters are applied in a cascading manner.

No Relations Filter. The sentence does not contain any named entities for the distinguished attributes.

Other Relations Filter. The sentence contains named entities that are not covered by the domain ontology.

Contextual Filter. The sentence contains *indexicals*, such as "I", "you", "that" or cohesive markers of rhetorical relations that connect it to some part of the preceding text, which means that the sentence cannot be interpreted out of context. These include discourse markers, such as list item markers with LS as the POS tag, that signal the organization structure of the text [Hirschberg and Litman 1987; Prasad et al. 2005], as well as discourse connectives that signal semantic and pragmatic relations of the sentence with other parts of the

Table I. Lexicalizations for Distinguished Attributes in the Restaurant Domain

Dist. Attr.	Lexicalization
food	food, meal
service	service, staff, waitstaff, wait staff, server, waiter, waitress
atmosphere	atmosphere, decor, ambience, decoration
value	value, price, overprice, pricey, expensive, inexpensive, cheap, affordable, afford
overall	recommend, place, experience, establishment

text [Knott 1996], such as coordinating conjunctions at the beginning of the utterance like "and" and "but" etc., and conjunct adverbs such as "however", "also", "then".

Unknown Words Filter. The sentence contains words not in WordNet [Fellbaum 1998] (which includes typographical errors), or POS tags contain NN (Noun), which may indicate an unknown named entity, or the sentence has more than a fixed length of words,[2] suggesting that its meaning cannot be estimated using only the occurrence of named entities.

Parsing Filter. The sentence fails the parsing to DSyntS conversion. Failures are automatically detected by comparing the original sentence with the one realized by RealPro taking the converted DSyntS as an input.

Duplicate Filter. The triple for the sentence has already been observed.

Finally, we add the triples that survive the filtering process to the dictionary.

3. EXPERIMENT

We first create generation dictionaries in the restaurant and hotel domains, and then evaluate the dictionaries using both objective and subjective criteria.

3.1 Creating Generation Dictionaries

3.1.1 *Restaurant Domain.* We collected user reviews from we8there.com (http://www.we8there.com/) and identified seven distinguished attributes from the ratings and the tabular data, namely, food, service, atmosphere, value, overall, location, and food type. Table I shows the lexicalizations of the distinguished attributes for the domain ontology in Section 2.2.

Out of 18,466 review sentences, we obtained 451 mappings; 2.4% of all sentences were used to create the mappings. Table II shows the number of sentences filtered and retained by each filter. Named entities for food subtypes (e.g., pizza, wine), person names, country names, dates (e.g., today, tomorrow, Aug. 26th) or prices (e.g., 12 dollars), or POS tag CD for numerals were used by the Other Relations Filter to detect relations not in the domain ontology.

3.1.2 *Hotel Domain.* We collected user reviews in the hotel domain from travelocity (http://dest.travelocity.com/Reviews/) and identified eight distinguished attributes: service, entertainment, sports (sports/activities), overall, facility (public facilities), room, dining, and location, where location is the only

[2]We experimentally derived 20 as a suitable threshold.

Table II. Filtering Statistics: The Number of Sentences Filtered and Retained by Each Filter in the Restaurant Domain

Filter	Filtered	Retained
Initial # of sentences	–	18,466
No Relations Filter	7,947	10,519
Other Relations Filter	5,351	5,168
Contextual Filter	2,973	2,195
Unknown Words Filter	1,467	728
Parsing Filter	216	512
Duplicates Filter	61	**451**

Table III. Lexicalizations for Distinguished Attributes in the Hotel Domain

Dist. Attr.	Lexicalizations
service	service, staff
entertainment	entertainment, attraction, amusement, fun, show
sports	sport, activity, workout, exercise, athletics
overall	hotel, overall, recommend, place, experience
facility	facility, amenity, appliance, equipment
room	room
dining	dining, dinner, food, meal

categorical attribute. From the attributes, we created the domain ontology:

$$\left\{\begin{array}{l}\text{HOTEL has servicequality}\\ \text{HOTEL has entertainmentquality}\\ \text{HOTEL has sportsquality}\\ \text{HOTEL has overallquality}\\ \text{HOTEL has facilityquality}\\ \text{HOTEL has roomquality}\\ \text{HOTEL has diningquality}\\ \text{HOTEL has location.}\end{array}\right.$$

Table III shows lexicalizations of the rating-related distinguished attributes. Exactly the same filters were used as those for the restaurant domain except that food subtypes were not detected by the named-entity tagger. Out of 20,723 review sentences, 536 mappings were obtained; 2.6% of all sentences were used to create the mappings. (See Table IV for the filtering statistics.)

3.2 Objective Evaluation

We evaluated the learned mappings with respect to domain coverage, linguistic variation and generativity.

3.2.1 *Domain Coverage.* For the mappings to be useful, they must have good domain coverage. Table V shows the distribution of the 327 mappings realizing a single scalar-valued relation categorized by the associated rating

Table IV. Filtering Statistics: The Number of Sentences Filtered and Retained by Each Filter in the Hotel Domain

Filter	Filtered	Retained
Initial # of sentences	–	20,723
No Relations Filter	12,413	8,580
Other Relations Filter	1,293	7,287
Contextual Filter	3,881	3,406
Unknown Words Filter	2,541	865
Parsing Filter	257	608
Duplicates Filter	72	**536**

Table V. Domain Coverage of Single Scalar-Valued Relation Mappings in the Restaurant Domain

Rating / Dist.Attr.	1	2	3	4	5	Total
food	5	8	6	18	57	94
service	15	3	6	17	56	97
atmosphere	0	3	3	8	31	45
value	0	0	1	8	12	21
overall	3	2	5	15	45	70
Total	23	15	21	64	201	327

score in the restaurant domain.[3] For example, there are 57 mappings with $\mathcal{R}$ of "RESTAURANT *has foodquality* = 5," and a large number of mappings for both the foodquality and servicequality relations. Although we could not obtain mappings for some relations, such as price = {1,2}, coverage for expressing a single relation is fairly complete; domain coverage is 88% (22/25). When we look into the corpus, we find that there are quite a few sentences associated with the ratings price = {1,2}, with 2,449 and 1,762 sentences, respectively. Therefore, the lack of mappings for these ratings suggests that reviewers rarely mention such relations or that our lexicalizations for value quality are not sufficient.

There are also mappings that express several relations. Table VI shows the counts of mappings for multirelation mappings with those containing a food or service relation occurring more frequently as in the single scalar-valued relation mappings. We found only 21 combinations of relations which is surprising given the large potential number of combinations. (There are 50 combinations if we treat relations with different scalar values differently.) We also find that most of the mappings have just two or three relations, perhaps suggesting that system utterances should not express too many relations in a single sentence.

In the hotel domain, we found 58 distinct semantic representation patterns, including 24 single-relation patterns (23 rating-related relations and one for location) and 16 multirelation patterns (34 combinations of relations when scalar values are considered). Table VII shows the distribution of single-relation mappings categorized by the associated rating score and Table VIII shows the

[3]There are two other single-relation but not scalar-valued mappings that concern LOCATION in our mappings.

Table VI. Counts for Multirelation Mappings in the Restaurant Domain

#	Combination of Dist. Attrs	Count
1	food-service	39
2	food-value	21
3	atmosphere-food	14
4	atmosphere-service	10
5	atmosphere-food-service	7
6	food-foodtype	4
7	atmosphere-food-value	4
8	location-overall	3
9	food-foodtype-value	3
10	food-service-value	2
11	food-foodtype-location	2
12	food-overall	2
13	atmosphere-foodtype	2
14	atmosphere-overall	2
15	service-value	1
16	overall-service	1
17	overall-value	1
18	foodtype-overall	1
19	food-foodtype-location-overall	1
20	atmosphere-food-service-value	1
21	atmosphere-food-overall-service-value	1
	Total	122

Table VII. Domain Coverage of Single Scalar-Valued Relation Mappings in the Hotel Domain

Rating / Dist.Attr.	1	2	3	4	5	Total
service	7	4	4	26	70	111
entertainment	1	0	0	0	0	1
sports	0	0	0	0	0	0
overall	22	14	20	72	84	212
facility	1	1	0	2	1	5
room	18	12	12	25	55	122
dining	1	0	0	4	3	8
Total	50	31	36	129	213	459

counts for multirelation mappings. (There are six other mappings that concern LOCATION.) It is noticeable that very few mappings for the entertainment and sports distinguished attributes were induced, making the domain coverage rather low at 63% (23/35), which means some additional manual mappings would be necessary for systems to be able to generate utterances in this domain. However, when we exclude entertainment/sports, the domain coverage is as good as that in the restaurant domain. An analysis of the corpus revealed that this sparseness comes from reviewers' tendency to mention subconcepts of entertainment/sports, such as casinos, pools, and beaches; rarely did they mention entertainment or sports quality as a whole.

Table VIII. Counts for Multirelation Mappings in the Hotel Domain

#	Combination of Dist. Attrs	Count
1	room-service	24
2	overall-service	20
3	location-overall	5
4	overall-room	4
5	overall-room-service	3
6	dining-service	3
7	facility-room	2
8	dining-room-service	2
9	service-sports	1
10	room-sports	1
11	location-service	1
12	facility-sports	1
13	facility-overall	1
14	dining-room	1
15	dining-overall	1
16	dining-facility-room-service	1
	Total	71

Table IX. Common Syntactic Patterns of DSyntSs in the Restaurant and Hotel Domains, Flattened to a POS Sequence for Readability (NN, VB, JJ, RB, CC stand for noun, verb, adjective, adverb, and conjunction, respectively)

Rank	Restaurant Domain		Hotel Domain	
	Syntactic Pattern	Ratio	Syntactic Pattern	Ratio
1	NN VB JJ	20.4%	NN VB JJ	15.9%
2	NN VB RB JJ	11.5%	NN VB RB JJ	9.5%
3	JJ NN	8.0%	JJ NN	7.6%
4	NN VB JJ CC JJ	5.5%	NN VB JJ CC JJ	5.9%
5	RB JJ NN	4.9%	NN VB RB JJ CC JJ	5.4%

3.2.2 *Linguistic Variation.* We also wish to assess whether the linguistic variation of the learned mappings is greater than what we could easily have generated with a handcrafted dictionary or with a handcrafted dictionary augmented with aggregation operators as in Stent et al. [2004]. Thus, we first categorized the mappings by the patterns of the DSyntSs. Table IX shows the five most common syntactic patterns in both domains, indicating that 25–30% of the learned patterns consist of the simple form "X is ADJ," where ADJ is an adjective, or "X is RB ADJ," where RB is a degree modifier. In fact, approximately half of the learned mappings (50.3% and 50.0% in the restaurant and hotel domains, respectively) could be generated from these basic patterns by applying a combination operator that coordinates multiple adjectives or coordinates predications over distinct attributes.

However, there are 137 syntactic patterns in the restaurant domain (97 with unique syntactic structures and 21 with two occurrences), and there are 142 syntactic patterns in the hotel domain (96 with unique syntactic structures and 9 with two occurrences). Figure 2 shows examples of learned mappings

- **Restaurant Domain**
  - —Very disappointing experience for the money charged. (overall=1, value=2)
  - —The food is excellent and plentiful at a reasonable price. (food=5, value=5)
  - —The food is exquisite as well as the service and setting. (food=5, service=5)
  - —The food was spectacular and so was the service. (food=5, foodtype, value=5)
  - —Best FOODTYPE food with a great value for money. (food=5, foodtype, value=5)
  - —This is the best place to eat FOODTYPE food in LOCATION. (food=5, foodtype)
  - —Simply amazing FOODTYPE food. (food=5, foodtype)
  - —RESTAURANTNAME is the best of the best for FOODTYPE food. (food=5)
  - —The food is to die for. (food=5)
  - —What incredible food. (food=5)
  - —Very pleasantly surprised by the food. (food=4)
  - —The food has gone downhill. (food=1)
  - —This is a quiet little place with great atmosphere. (atmosphere=5, overall=5)
- **Hotel Domain**
  - —Great service by a friendly staff. (service=5)
  - —The staff was unfriendly, the rooms were poorly kept and worn. (room=1, service=1)
  - —The rooms were dirty old and smelled. (room=1)
  - —The hotel is fairly new and so the rooms are nice and clean. (overall=5, room=5)
  - —Great hotel to stay in. (overall=5)
  - —Great place for an affordable place to sleep. (overall=4)
  - —The rooms are just average hotel rooms. (overall=2, room=3)
  - —Just a place to sleep. (overall=2)
  - —This hotel looks old and is old. (overall=2)
  - —This is the best hotel to stay in LOCATION. (location, overall=5)
  - —The entire facility was extremely dated. (facility=2)
  - —There was NO entertainment. (entertainment=1)
  - —Food was great along with service. (dining=5, service=4)

Fig. 2. Acquired generation patterns (with shorthand for relations in brackets) whose syntactic patterns occurred only once.

with distinct syntactic structures. It would be surprising to see this type of variety in a handcrafted generation dictionary. In addition, the learned mappings contain 275 distinct lexemes with a minimum of 2, maximum of 15, and mean of 4.63 lexemes per DSyntS in the restaurant domain, and 254 distinct lexemes, with a minimum of 2, maximum of 15, and mean of 4.56 lexemes per DSyntS in the hotel domain, indicating that the method extracts a wide variety of expressions of varying lengths. When we merge the mappings of the two domains, we find 229 distinct syntactic patterns and 406 distinct lexemes, which shows the relative syntactical and lexical differences between the two domains.

Another interesting aspect of the learned mappings is the wide variety of adjectival phrases (APs). Tables X and XI show the APs in single scalar-valued relation mappings for *food* and *room* categorized by the associated ratings in the restaurant domain. Tables for other scalar-valued attributes can be found in the appendix. The bold/italicized fonts indicate whether the adjectives are registered in the General Inquirer lexicon (http://www.wjh.harvard.edu/~inquirer/) as positive (bold) or negative (italicized) words. Since the General Inquirer lexicon covers domain-independent positive/negative adjectives, adjectives not in bold/italicized font can be considered domain-specific ones. Since we observe

Table X. Adjectival Phrases (APs) in Single Scalar-Valued Relation Mappings for *foodquality*

food = 1	burnt, very ordinary, *awful*, *bad*, *cold*
food = 2	flavored, not enough, very **good**, very *bland*, **acceptable**, *bad*
food = 3	flavorful but *cold*, rather *bland*, very **good**, **adequate**, **pretty good**, *bland* and *mediocre*
food = 4	absolutely **wonderful**, awesome, rather **good**, really **good**, very very **good**, very **fresh** and tasty, very **good**, **decent**, **excellent**, **good**, **good** and **generous**, **great**, **outstanding**, **traditional**
food = 5	absolutely delicious, absolutely **fantastic**, absolutely **great**, absolutely **terrific**, awesome, delectable and **plentiful**, delicious, delicious but simple, large and satisfying, quick and *hot*, simply **great**, so delicious, so very tasty, superb, very **good**, **ample, well** seasoned and *hot*, **best**, **excellent**, **exquisite**, **fabulous**, **fancy** but tasty, **fantastic**, **fresh**, **good**, **great**, **just fantastic**, **outstanding**, **plentiful** and tasty, **plentiful** and **outstanding**, **terrific**, **tremendous**, **wonderful**, *hot*, *incredible*

Table XI. Adjectival Phrases (APs) in Single Scalar-Valued Relation Mappings for *Roomquality*

room = 1	disgusting, gross and *nasty*, old, old and *dirty*, particularly **clean** or **comfortable**, really gross, small, very *dirty*, very *nasty*, whole, *awful*, *dirty*, *horrible*, *nasty*, *terrible*
room = 2	average, old, old and *dirty*, outdated and *dirty*, really moldy and *stale*, really old and ragged, very **clean**, *stale*
room = 3	average, very quiet, very **basic**, very **clean**, **comfortable**, **decent**, **nice**, *bland*
room = 4	huge, large, new, public, so **clean**, spacious, understated and **warm**, very **clean**, very **clean** and **pleasant**, very **comfortable**, very **nice**, **bright**, **clean, comfortable** and spacious, **clean**, **clean** and **comfortable**, **clean** and **neat**, **clean** and **nice**, **great**, **perfect**, **spotless**
room = 5	beautiful, beautiful, **clean** and spacious, beautiful, **spotless** and **comfortable**, extremely large, huge, impeccable, large, large and very **comfortable**, large, **clean** and quiet, new, newly renovated, quite spacious, renovated, so spacious and **luxurious**, spacious, spacious, **comfortable** and **well** appointed, very large, very nicely decorated, very spacious, very spacious and beautiful, very **clean, bright** and **attractive**, very **clean**, very **clean** and **comfortable**, very **clean** and **nice**, very **comfortable**, very **cozy**, very **nice**, **cheerful** and **unique**, **clean**, **clean** and very **nice**, **clean** and very **reasonable**, **clean** and **comfortable**, **clean** and **nice**, **comfortable**, **elegant**, **extraordinary**, **fabulous**, **fantastic**, **great**, **nice**, **nice** and spacious, **spotless**, **wonderful**

many such adjectives, this indicates that our approach can make systems speak with a vocabulary suited to the domain.

Moreover, the meanings for some of the learned APs are very specific to the particular attribute, for example, "cold" and "burnt" associated with foodquality of 1, "attentive and prompt" for servicequality of 5, "silly and inattentive" for servicequality of 1, and "mellow" for atmosphere of 5. In addition, our method places the APs on a more fine-grained scale of 1 to 5, similar to the strength classifications in Wilson et al. [2004], in contrast to other automatic methods that classify expressions into a binary positive or negative polarity (e.g., Turney [2002]).

• **Content Plan**
—RST relations: justify(p1, p2), justify(p1, p3), justify(p1, p4)
—propositions:
p1. assert-best(Babbo) ⇔ <no corresponding semantic representation>
p2. assert-food_quality(Babbo, superb) ⇔ food=5
p3. assert-service(Babbo, excellent) ⇔ service=5
p4. assert-decor(Babbo, superb) ⇔ atmosphere=5
Original SPaRKy utterance: Babbo has excellent decor and superb food quality with excellent service. It has the best overall quality among the selected restaurants.
With learned mappings:
u1. Babbo has superb food quality, the service is exceptional and the atmosphere is very creative. It has the best overall quality among the selected restaurants.
u2. The food is phenomenal and the atmosphere is very unique. Babbo has excellent service. It has the best overall quality among the selected restaurants.
u3. Great food and small cozy atmosphere. Babbo has excellent service. It has the best overall quality among the selected restaurants.
u4. Since the service is fast and friendly, the food is really good and Babbo has excellent decor, it has the best overall quality among the selected restaurants.

Fig. 3. Example utterances for the restaurant domain incorporating learned DSyntSs in SPaRKy. The symbol ⇔ indicates how a SPaRKy proposition is mapped to a semantic representation (in shorthand). The infer relation is equal to the joint relation in RST.

3.2.3 *Generativity.* Our motivation for deriving syntactic representations for the learned expressions was to explore the possibility of using an off-the-shelf sentence planner to derive new combinations of relations and apply aggregation and other syntactic transformations. We examined how many of the learned DSyntSs can be combined with each other by taking every pair of DSyntSs in the mappings and applying the built-in merge operation in the SPaRKy generator [Stent et al. 2004]. We found that only 306 combinations out of a potential 81,318 (0.37%) and 544 combinations out of a potential 138,422 (0.39%) were successful in the restaurant and hotel domains, respectively. This is because the merge operation in SPaRKy requires that the subjects and the verbs of the two DSyntSs be identical, for instance, the subject is RESTAURANT (or HOTEL) and the verb is "has", whereas the learned DSyntSs often place the attribute in the subject position as a definite noun phrase.

However, the learned DSyntS can be incorporated into SPaRKy. The SPaRKy generator takes as input a set of propositions and the rhetorical relations among them. Each proposition has a corresponding DSyntS (assigned by hand), and the DSyntSs for propositions are aggregated by Rhetorical Structure Theory (RST) transformation rules to create a sentence-plan tree satisfying the relations. We can substitute the DSyntSs for the propositions with the ones in the learned mappings in the aggregation process. We prepared several rules to map SPaRKy propositions to our semantic representations to enable the substitution. Figure 3 and Figure 4 show how an utterance would change by incorporating the learned mappings in each domain. The handcrafted mappings used for comparison as the original SPaRKy utterance are also used in the subjective evaluation (see Section 3.3). The resulting utterances seem more natural and colloquial; the subjective evaluation examines whether this is true.

• **Content Plan**
—RST relations: infer(p1, p2, p3, p4, p5, p6, p7, p8)
—propositions:
p1. assert-service(Bellagio, excellent) ⇔ service=5
p2. assert-entertainment(Bellagio, excellent) ⇔ entertainment=5
p3. assert-sports(Bellagio, excellent) ⇔ sports=5
p4. assert-overall(Bellagio, excellent) ⇔ overall=5
p5. assert-facility(Bellagio, excellent) ⇔ facility=5
p6. assert-room(Bellagio, excellent) ⇔ room=5
p7. assert-dining(Bellagio, excellent) ⇔ dining=5
p8. assert-nbhd(Bellagio, Las Vegas) ⇔ location

Original SPaRKy utterance: Bellagio is located in Las Vegas and has excellent dining quality. It is excellent for sports and activities. Bellagio, which has excellent rooms and excellent overall quality, has excellent facilities. It has excellent entertainment quality. It has excellent service.

With learned mappings:

u1. Really all fairly good for Las Vegas, Bellagio is excellent for sports and activities and the staff is super friendly and extremely attentive. Bellagio has excellent entertainment quality and the facility is clean and comfortable. Great food and the room is very large. The hotel is magical.

u2. Very nicely decorated rooms. Not bad for Las Vegas. Bellagio is excellent for sports and activities and the facility is clean and comfortable. Regardless, great service, the food is excellent and the hotel is great. Bellagio has excellent entertainment quality.

u3. The room is great. The staff is helpful and courteous and Bellagio has excellent entertainment quality. Las Vegas is just beautiful and the facility is clean and comfortable. The hotel is wonderful, Bellagio is excellent for sports and activities and the food is excellent.

u4. Las Vegas is just beautiful and the whole experience is awesome. Very nicely decorated rooms and Bellagio has excellent facilities. The food is delicious. Bellagio is excellent for sports and activities, the staff is very pleasant and accommodating and has excellent entertainment quality.

Fig. 4. Example utterances incorporating learned DSyntSs in SPaRKy. The symbol ⇔ indicates how a SPaRKy proposition is mapped to our semantic representation (in shorthand). The infer relation is equal to the joint relation in RST.

3.3 Subjective Evaluation

We also wish to evaluate the learned dictionary in the context of our application: the generation of evaluative utterances (recommendations and comparisons) in the restaurant and hotel domains. A primary concern is whether the precision of our procedure for deriving semantic representations is high enough. A secondary concern is whether the mappings are natural and appropriate for spoken language generation. Thus, we evaluate the obtained mappings in two respects: (i) the consistency between the semantic representation and the realization and (ii) the naturalness of the realization. In the evaluation, we compare handcrafted mappings (a baseline) with the learned mappings. Figure 3 and Figure 4 show sample recommendations in both the restaurant and hotel domains. Each figure shows the content plan with the propositions realized, an original SPaRKy utterance (one of several) and sample utterances incorporating the learned mappings for the same content plan.

In the restaurant domain, the baseline mappings were those from Stent et al. [2004] except that we changed the word *decor* to *atmosphere* and added

service	⇔	HOTELNAME has ADJ service.
entertainment	⇔	HOTELNAME has ADJ entertainment quality.
sports	⇔	HOTELNAME is ADJ for sports and activities.
overall	⇔	HOTELNAME has ADJ overall quality.
facility	⇔	HOTELNAME has ADJ facilities.
room	⇔	HOTELNAME has ADJ rooms.
dining	⇔	HOTELNAME has ADJ dining quality.
location	⇔	HOTELNAME is located in LOCATION.

Fig. 5. Handcrafted mappings in the hotel domain. ADJ is "mediocre", "decent", "good", "very good", or "excellent" depending on the rating values 1–5.

five mappings for *overall*. For scalar relations, this consists of the realization "RESTAURANT *has* ADJ LEX," where ADJ is "mediocre", "decent", "good", "very good", or "excellent" for rating values 1–5 and LEX is "food quality", "service", "atmosphere", "value", or "overall" depending on the relation. RESTAURANT is filled with the name of a restaurant at runtime. For example, 'RESTAURANT *has foodquality = 1*' is realized as "RESTAURANT *has mediocre food quality.*" The location and food type relations are mapped to "RESTAURANT *is located in* LOCATION" and "RESTAURANT *is a* FOODTYPE *restaurant.*" In the hotel domain, we created new handcrafted mappings (Figure 5).

In the restaurant domain, the learned mappings include 23 distinct semantic representations for a single-relation (22 for scalar-valued relations and one for location) and 50 for multirelations. Therefore, using the handcrafted mappings, we first created 23 utterances for the single relations. We then created three utterances for each of 50 multirelations using different clause-combining operations from Stent et al. [2004]. This gave a total of 173 baseline utterances which, together with 451 learned mappings, yielded 624 utterances for evaluation. In the same way, we created 662 [536 (number of learned mappings) + 24 (number of single-relation mappings) + 34 (number of multirelation mappings) × 3] utterances in the hotel domain.

Ten subjects and five subjects, all native English speakers, evaluated the mappings in the restaurant and hotel domains, respectively. They were presented with pairs comprising a semantic representation (in shorthand) and a realization (utterance) for all learned and baseline mappings which were randomly ordered. The subjects were asked to express their degree of agreement, on a scale of 1 (lowest) to 5 (highest), with the statement (a) The meaning of the utterance is consistent with the ratings expressing their semantics and with the statement (b) The style of the utterance is very natural and colloquial. They were asked not to correct their decisions and also to rate each utterance on its own merit.

3.3.1 *Results.* Tables XII and XIII show the means and standard deviations of the scores for baseline vs. learned utterances for consistency and naturalness in the restaurant and hotel domains, respectively. In both domains, a t-test shows that the consistency of the learned mappings is significantly lower than the baseline and that their naturalness is significantly higher than the baseline.

Table XII. Consistency and Naturalness Scores Averaged over 10 Subjects in the Restaurant Domain

	Baseline		Learned		Statistical Significance by Welch's t-test
	Mean	sd.	Mean	sd.	
Consistency	**4.714**	0.588	**4.459**	0.890	+ (df = 4712, p < .001)
Naturalness	**4.227**	0.852	**4.613**	0.844	+ (df = 3107, p < .001)

Table XIII. Consistency and Naturalness Scores Averaged Over Five Subjects in the Hotel Domain

	Baseline		Learned		Statistical Significance by Welch's t-test.
	Mean	sd.	Mean	sd.	
Consistency	**4.533**	0.906	**4.176**	0.942	+ (df = 975, p < .001)
Naturalness	**3.603**	1.062	**3.948**	1.108	+ (df = 977, p < .001)

Table XIV. Distribution of Mean Consistency/Naturalness Scores of the Learned Mappings

Mean Score (s) Range	Restaurant Domain		Hotel Domain	
	cons.	nat.	cons.	nat.
$1.0 \leq s < 1.5$	2	2	2	3
$1.5 \leq s < 2.0$	3	1	5	8
$2.0 \leq s < 2.5$	2	3	11	15
$2.5 \leq s < 3.0$	7	1	16	28
$3.0 \leq s < 3.5$	21	15	53	73
$3.5 \leq s < 4.0$	41	24	59	77
$4.0 \leq s < 4.5$	92	62	174	197
$4.5 \leq s \leq 5.0$	283	343	216	135

However, consistency is still high. Table XIV shows the distribution of mean consistency/naturalness scores of the learned mappings which indicates that only a very small proportion of the learned utterances have a mean consistency/naturalness score lower than 3. Although there is some fluctuation in the quality of mappings across domains, by and large, the human judges felt that the inferred semantic representations were consistent with the meaning of the learned expressions and that they were natural. The correlation coefficient between consistency and naturalness scores is 0.42 and 0.33 in the restaurant and hotel domain, respectively, indicating that consistency is not strongly related to naturalness.

We also performed an ANOVA (ANalysis Of VAriance) of the effect of each relation in $\mathcal{R}$ on naturalness and consistency. In the restaurant domain, there were no significant effects except that mappings combining *food*, *service*, and *atmosphere* were significantly worse for naturalness (df = 1, F = 7.79, p = 0.005). However, there is a tendency for mappings to be rated higher for the *food* attribute (df = 1, F = 3.14, p = 0.08) and the *value* attribute (df = 1, F = 3.55, p = 0.06) for consistency.

In the hotel domain, mappings expressing *overall* were significantly worse for naturalness (df = 1, F = 7.32, p = 0.007). For consistency, attributes *sports* (df = 1, F = 13.15, p = 0.0003), *room* (df = 1, F = 4.31, p = 0.04), and combinations of

service and *room* ($df = 1$, $F = 4.19$, $p = 0.04$), and *overall* and *dining* ($df = 1$, $F = 6.24$, $p = 0.004$) had negative effects. On the other hand, mappings expressing *service* ($df = 1$, $F = 10.85$, $p = 0.001$), and those combining *service*, *overall* and *room* ($df = 1$, $F = 8.23$, $p = 0.004$) were rated significantly higher. By and large, some attributes seem to be more difficult to learn than others, perhaps because of the difference in the quality of lexicalizations of distinguished attributes and the choice of ratings of the review websites.

4. CONCLUSION AND FUTURE WORK

We proposed automatically obtaining mappings between semantic representations and realizations from reviews with individual ratings. Experimental results show that (1) the learned mappings provide good coverage of the domain ontology and exhibit good linguistic variation; (2) the consistency between the semantic representations and realizations is high; (3) the naturalness of the realizations is significantly higher than the baseline; and (4) although there are some differences in the quality of mappings across domains, the method is likely to be applicable to multiple domains.

One limitation of the work is with the evaluation method. The evaluation in the restaurant domain was based on a comparison with a previously developed generation dictionary which had been evaluated independently and shown to produce high quality output [Stent et al. 2004]. However, we had no such preexisting dictionary in the hotel domain, and therefore created it ourselves for the purposes of evaluation. Thus one limitation of the subjective evaluation is that there could be better, more natural realizations in the hotel domain which we did not explore.

There are also limitations of our method. Even though consistency is rated highly by human subjects, this is a judgment of whether the polarity of the learned mapping is correctly placed on the 1 to 5 rating scale. Thus, alternate ways of expressing, for example *foodquality* = 5, shown in Table X, cannot be guaranteed to be synonymous, which is a requirement for use in spoken language generation. An examination of the adjectival phrases in Table X shows that different aspects of the food are discussed. For example, "ample" and "plentiful" refer to the portion size, "fancy" may refer to the presentation, and "delicious" describes flavors. One solution would be to automatically extend the ontology to represent these subattributes of the food attribute and subattributes in general. Another solution would be to use frequency information or other information to find the most general terms which are likely to refer to overall food quality rather than a subattribute [Lapata and Keller 2005]. We believe this could be done automatically in future work. We have made no use of frequency information as yet, but there is much related work that bases judgments of syntax/semantics mapping on frequency information [Turney 2002; Soderland 2007; Etzioni et al. 2005].

Another problem with consistency is that the same AP, for example, "very good" in Table X, may appear with multiple ratings. For instance, "very good" is used for every foodquality rating from 2 to 5. This may reflect a difference in the use of language by individuals, related to idiolect or personality [Reiter

and Sripada 2002; Mairesse and Walker 2007]. One solution to this would be to use frequency information to decide which rating was the most accurate or to eliminate adjectives that appear with more than one rating. However, even without further extensions, the method presented here could reduce the amount of time a system designer spends developing the spoken language generator and could increase the naturalness of spoken language generation.

Another issue is that the recall appears to be quite low given that all of the sentences concern the same domain: only around 2.5% of the sentences could be used to create the mappings. One way to increase recall might be to automatically augment the list of distinguished attribute lexicalizations, using WordNet or work on automatic identification of synonyms such as Lin and Pantel [2001]. However, the method has high precision, and automatic techniques may introduce noise.

In addition, recall is greatly reduced by the parsing/generation filter which verifies that the automatically derived syntactic representation (DSyntS) can be used to regenerate the exact string. Thus the utility of syntactic structures in the mappings should be further examined, especially given the failures in DSyntS conversion. An alternative would be to leave some sentences unparsed and use them as templates with hybrid generation techniques [White and Caldwell 1998].

A related recall issue is that the filters are in some cases too strict. For example, the contextual filter is based on POS-tags so that sentences that do not require the prior context for their interpretation are eliminated such as those containing subordinating conjunctions like “because”, “when”, “if”, whose arguments are both given in the same sentence [Prasad et al. 2005]. In addition, recall is affected by the domain ontology, and the automatically constructed domain ontology from the review Web pages may not completely cover the domain. In some review domains, the attributes that get individual ratings are a limited subset of the domain ontology. Techniques for automatic feature identification [Hu and Liu 2005; Popescu and Etzioni 2005] could help here although these techniques currently do not automatically identify different lexicalizations of the same feature.

A different type of limitation is that dialogue systems need to generate utterances for information gathering, whereas the mappings we obtained can only be used for information presentation. Thus, mappings for this purpose would have to be constructed by hand as in current practice, or perhaps other types of corpora or resources could be utilized.

Finally, the comparison of evaluation results in the two domains suggested that some distinguished attributes are not necessarily expressed by their straightforward lexicalizations: sports and entertainment are usually expressed by their subconcepts, making the domain coverage low. This provides another motivation for considering how to automatically extend the domain ontology to better suit the domain. The difference in the quality of the mappings between the domains may also indicate a need for further refinement of the method. Although we did not find outstanding faults with the mappings in the hotel domain, we need to investigate why they received relatively low scores.

Despite the limitations, the results of our experiments suggest that our approach is promising. We would like to pursue ways to overcome the limitations to further facilitate the development of a generation module of spoken dialogue systems.

APPENDIXES

A. ADJECTIVAL PHRASES (APS) IN THE RESTAURANT DOMAIN

service = 1	horrendous, inattentive, forgetful and slow, really slow, still *marginal*, young, **great**, *awful*, *bad*, *horrible*, *marginal*, *silly* and inattentive, *terrible*
service = 2	overly slow, very slow and inattentive
service = 3	very **friendly**, **friendly** and knowledgeable, **good**, **pleasant**, **prompt**, *bad*, *bland* and *mediocre*
service = 4	extremely **friendly** and **good**, extremely **pleasant**, really **friendly**, so **nice**, swift and **friendly**, very **friendly**, very **friendly** and accommodating, **attentive**, **fantastic**, **friendly**, **friendly** and **helpful**, **good**, **great**, **great** and **courteous**, **prompt** and **friendly**, *all* very **warm** and welcoming
service = 5	extremely **friendly**, impeccable, intrusive, legendary, quick and **cheerful**, superb, the most **attentive**, very timely, very **attentive**, very **congenial**, very **courteous**, very **friendly**, very **friendly** and totally personal, very **friendly** and welcoming, very **friendly** and **helpful**, very **friendly** and **pleasant**, very **good**, very **helpful**, **excellent**, **excellent** and **friendly**, **fabulous**, **fantastic**, **friendly**, **friendly** and very **attentive**, **friendly** and **helpful**, **good**, **great, prompt** and **courteous**, **great**, **happy** and **friendly**, **outstanding**, **pleasant**, **polite, attentive** and **prompt**, **prompt** and **courteous**, **prompt** and **pleasant**, **stupendous**, **warm** and **friendly**, **wonderful**, *all* **courteous**, *unbelievable*
atmosphere = 2	eclectic, **unique** and **pleasant**
atmosphere = 3	busy, **pleasant** but extremely *hot*
atmosphere = 4	quite **nice** and simple, typical, very trendy, very **casual**, **fantastic**, **great**, **wonderful**
atmosphere = 5	beautiful, interior, phenomenal, quite **pleasant**, unbelievably beautiful, very relaxing, very **comfortable**, very **cozy**, very **friendly**, very **intimate**, very **nice**, very **nice** and relaxing, very **pleasant**, **comfortable**, **excellent**, **great**, **lovely**, **mellow**, **nice**, **nice** and **comfortable**, **pleasant**, **warm** and contemporary, **warm** and very **comfortable**, **wonderful**
value = 3	very **reasonable**
value = 4	very **good**, **great**, **pretty good**, **reasonable**
value = 5	extremely **reasonable**, totally **reasonable**, very **good**, very **reasonable**, **best**, **good**, **great**, **reasonable**
overall = 1	thoroughly humiliating, **just** *bad*, **nice**
overall = 2	really *bad*, **great**
overall = 3	interesting, really **fancy**, **decent**, **great**, *bad*
overall = 4	never busy, not very busy, recommended, **excellent**, **good**, **great**, **just great**, **outstanding**, **wonderful**
overall = 5	awesome, capacious, extremely **pleasant**, local, new, overall, overwhelmingly **pleasant**, pampering, really **great**, really **neat**, really **nice**, really *cool*, tasty, truly **great**, ultimate, very **enjoyable**, very **excellent**, very **good**, very **nice**, very **wonderful**, **amazing**, **delightful**, **fantastic**, **good**, **great**, **marvelous**, **neat**, **peaceful** but idyllic, **special**, **unique** and **enjoyable**, **warm** and **friendly**, **wonderful**

B. ADJECTIVAL PHRASES (APS) IN THE HOTEL DOMAIN

service = 1	clueless and not **helpful**, very *poor*, very *unpleasant* and *rude*, **friendly**, *awful*, *rude*, *rude* and very uncooperative
service = 2	uncooperative and unprofessional, very **nice** and apologetic, **pretty** *poor*, *all* other, *poor*
service = 3	very **friendly** and **helpful**, very **helpful** and **friendly**, **friendly**, **nice**
service = 4	extremely **courteous** and **efficient**, extremely **friendly** and **helpful**, so **friendly**, very **friendly**, very **friendly** and **attentive**, very **helpful**, very **helpful** and **friendly**, very **helpful** and **wonderful**, very **polite** and **helpful**, **excellent**, **friendly**, **friendly** and knowledgeable, **friendly** and very **helpful**, **friendly** and **helpful**, **good**, **great**, **helpful** and **friendly**, **pleasant**, **super friendly**
service = 5	awesome, entire, exceptional and very **helpful**, exceptionally **polite** and **courteous**, extremely **attentive**, extremely **friendly**, extremely **friendly** and **helpful**, impeccable, phenomenal, very **attentive** and exceptionally **helpful**, professional and very **friendly**, really **friendly** and **helpful**, really **helpful**, so professional, so **courteous**, so **nice**, top, unbeatable, unbelievably **courteous** and **helpful**, very **efficient** and **helpful**, very **friendly, helpful** and professional, very **friendly**, very **friendly** and very **helpful**, very **friendly** and **courteous**, very **friendly** and **helpful**, very **helpful**, very **helpful** and **friendly**, very **helpful** and **prompt**, very **pleasant** and accommodating, **amazing**, **excellent**, **free**, **friendly,** personable and accommodating, **friendly**, **friendly** and accommodating, **friendly** and very **wonderful**, **friendly** and **helpful**, **great**, **great** and **helpful**, **helpful** and **courteous**, **helpful** and **friendly**, **marvelous**, **nice** and **helpful**, **outstanding**, **super friendly** and extremely **attentive**, **super nice** and **helpful**, **terrific**, **wonderful**, **wonderful** and **friendly**, **wonderful** and **helpful**, *all* very **friendly** and **helpful**, *all* very **nice**
overall = 1	absolutely disgusting, disgusting, old, old and outdated, old and *dirty*, really *bad*, so *bad*, very old, very old and *dirty*, very unprofessional, very *marginal*, whole, **pleasant**, *awful*, *bad*, *horrible*, *nasty*, *terrible*
overall = 2	old, really old and rickety, very seedy, very **clean**, very **comfortable**, *bad*, *cheap*, *terrible*
overall = 3	otherwise **clean**, very **comfortable**, very **nice**, very **satisfactory**, **best**, **clean**, **decent**, **fair**, **great**, **nice**, **nice** and quiet, **pretty good** and **inexpensive**, **pretty nice**, *mediocre*
overall = 4	affordable, beautiful, convenient, extremely affordable, extremely **clean**, fairly new, new, really beautiful, really **nice**, very affordable, very spacious, very **clean, comfortable** and beautiful, very **clean, comfortable** and quiet, very **clean**, very **clean** and quiet, very **clean** and **nice**, very **good**, very **nice**, very **pleasant**, **clean** and affordable, **clean** and very **nice**, **clean** and **luxurious**, **comfortable**, **decent**, **excellent**, **fabulous**, **good**, **great**, **impressive**, **lovely**, **nice, clean** and affordable, **nice**, **nice** and **friendly**, **perfect**, **pleasant**, **pretty nice**, **spotless**, **wonderful**, *bad*, *cool*
overall = 5	absolutely beautiful, absolutely **brilliant**, absolutely **remarkable**, awesome, beautiful, fairly new, large and **nice**, little, new, really **great**, simply the most beautiful, so **great**, somewhat pricey, truly **enjoyable**, very very **nice**, very **clean**, very **clean** and **comfortable**, very **good**, very **nice**, very **positive**, whole, **clean,** quiet and affordable, **clean** and **accessible**, **excellent**, **friendly**, **good**, **great**, **lovely**, **magical**, **nice**, **nice** and quiet, **perfect**, **real nice**, **wonderful**
facility = 1	broken
facility = 2	entire
facility = 4	quite **nice**, very **nice**
facility = 5	**clean** and **comfortable**
dining = 1	*expensive*
dining = 4	very **good**, **good**, **great**
dining = 5	delicious, **excellent**, **great**

REFERENCES

BAPTIST, L. AND SENEFF, S. 2000. GENESIS-II: A versatile system for language generation in conversational system applications. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP).* Vol. 3. 271–274.

Barzilay, R. and Lapata, M. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. 331–338.

Barzilay, R. and Lapata, M. 2006. Aggregation via set partitioning for natural language generation. In *Proceedings of Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. 359–366.

Barzilay, R. and Lee, L. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 164–171.

Barzilay, R. and Lee, L. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. 16–23.

Barzilay, R. and McKeown, K. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th ACL*. 50–57.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. 1977. GUS, a frame driven dialog system. *Artif. Intel. 8*, 155–173.

Brill, E. 1992. A simple rule-based part-ofspeech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP)*. 152–155.

Chambers, N. and Allen, J. 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialog*. 9–18.

Chen, J., Bangalore, S., Rambow, O., and Walker, M. 2002. Towards automatic generation of natural language generation systems. In *Proceedings of 19th COLING*. 1–7.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., , and Yates, A. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intel. 165*, 1, 91–134.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Feng, J., Reddy, S., and Sara clar, M. 2005. Webtalk: Mining websites for interactively answering questions. In *Proceedings of the 9th European Conference on Speech Communication and Technology*. 2485–2488.

Gildea, D. and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Comp. Ling. 28*, 3, 245–288.

Goddeau, D., Meng, H., Polifroni, J., Seneff, S., and Busayapongchai, S. 1996. A form-based dialogue manager for spoken language applications. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Vol. 2. 701–704.

Heisterkamp, P. 2001. Linguatronic: Product-level speech system for mercedes-benz car. In *Proceedings of Human Language Technology Conference (HLT)*. 1–2.

Higashinaka, R., Prasad, R., and Walker, M. 2005. Augmenting variation of system utterances using corpora in spoken dialogue systems. In *Proceedings of 2005 IEEE Automatic Speech Recognition and Understanding Workshop*. 262–267.

Higashinaka, R., Prasad, R., and Walker, M. 2006. Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In *Proceedings of COLING/ACL*. 265–272.

Higashinaka, R., Sudoh, K., and Nakano, M. 2006. Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems. *Speech Comm. 48*, 3–4, 417–436.

Hirschberg, J. and Litman, D. J. 1987. Now let's talk about NOW: Identifying cue phrases intonationally. In *Proceedings of 25th ACL.*. 163–171.

Hu, M. and Liu, B. 2005. Mining and summarizing customer reviews. In *Proc. KDD*. 168–177.

Kaji, N. and Kitsuregawa, M. 2006. Automatic construction of polarity-tagged corpus from html documents. In *Proceedings of COLING/ACL*. 452–459.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 423–430.

KNOTT, A. 1996. A Data-Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh, Edinburgh.

LANGKILDE-GEARY, I. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the International Natural Language Generation Conference (INLG)*. 17–24.

LAPATA, M. AND KELLER, F. 2005. Web-based models for natural language processing. *ACM Trans. Speech Lang. Proc. 2*, 1, 1–31.

LAVOIE, B. AND RAMBOW, O. 1997. A fast and portable realizer for text generation systems. In *Proceedings of 5th Conference on Applied Natural Language Processing*. 265–268.

LIN, D. 1998. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

LIN, D. AND PANTEL, P. 2001. Discovery of inference rules for question answering. *Natur. Lang. Engin. 7*, 4, 343–360.

MAIRESSE, F. AND WALKER, M. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th ACL*.

MELČUK, I. A. 1988. *Dependency Syntax: Theory and Practice*. SUNY, Albany, NY.

MOORE, J. D., FOSTER, M. E., LEMON, O., AND WHITE, M. 2004. Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*.

OH, A. AND RUDNICKY, A. 2000. Stochastic language generation for spoken dialogue systems. In *Proceedings of the NAACL-ANLP Workshop on Conversational Systems*. 27–32.

PANTEL, P. AND RAVICHANDRAN, D. 2004. Automatically labeling semantic classes. In *Proceedings of Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. 321–328.

PIERACCINI, R. AND LUBENSKY, D. 2005. Spoken language communication with machines: The long and winding road from research to business. In *Proceedings of IEA/AIE*. 6–15.

POPESCU, A.-M. AND ETZIONI, O. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. 339–346.

PRASAD, R., JOSHI, A., DINESH, N., LEE, A., MILTSAKAKI, E., AND WEBBER, B. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*.

RAMBOW, O., ROGATI, M., AND WALKER, M. 2001. Evaluating a trainable sentence planner for a spoken dialogue travel system. In *Proceedings of the 39th ACL*. 426–433.

REITER, E. AND DALE, R. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

REITER, E. AND SRIPADA, S. 2002. Human variation and lexical choice. *Comp. Ling. 28*, 545–553.

REITER, E., SRIPADA, S., AND ROBERTSON, R. 2003. Acquiring correct knowledge for natural language generation. *J. Artif. Intel. Resear. 18*, 491–516.

SENEFF, S. AND POLIFRONI, J. 2000. Formal and natural language generation in the mercury conversational system. In *Proceedings of the International Conference on Spoken Language Processing. (ICSLP)*. Vol. 2. 767–770.

SNYDER, B. AND BARZILAY, R. 2007. Database-text alignment via structured multilabel classification. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*. 1713–1718.

SODERLAND, S. 2007. Moving from textual relations to ontologized relations. In *Proceedings of the AAAI Spring Symposium on Machine Reading*.

STENT, A., PRASAD, R., AND WALKER, M. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd ACL*. 79–86.

THEUNE, M. 2003. From monologue to dialogue: natural language generation in OVIS. In *AAAI Spring Symposium on Natural Language Generation in Written and Spoken Dialogue*. 141–150.

TURNEY, P. D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th ACL*. 417–424.

WALKER, M., RAMBOW, O., AND ROGATI, M. 2002. Training a sentence planner for spoken dialogue using boosting. *Comput. Speech Lang. 16*, 3-4.

WHITE, M. AND CALDWELL, T. 1998. EXEMPLARS: A practical, extensible framework for dynamic text generation. In *Proceedings of the International Natural Language Generation Conference (INLG)*. 266–275.

WILSON, T., WIEBE, J., AND HWA, R. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of AAAI*. 761–769.

Received September 2006; revised May 2007; accepted June 2007 by Eiichiro Sumita