Learning to Generate Naturalistic Utterances Using Reviews in Spoken Dialogue Systems

Ryuichiro Higashinaka
NTT Corporation
rh@cslab.kecl.ntt.co.jp

Rashmi Prasad
University of Pennsylvania
rjprasad@linc.cis.upenn.edu

Marilyn A. Walker
University of Sheffield
walker@dcs.shef.ac.uk

Abstract

Spoken language generation for dialogue systems requires a dictionary of mappings between semantic representations of concepts the system wants to express and 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 on restaurant reviews. We test the hypothesis that user reviews that provide individual ratings for distinguished attributes of the domain entity make it possible to map review sentences to their semantic representation with high precision. Experimental analyses show that the mappings learned cover most of the domain ontology, and provide good linguistic variation. A subjective user evaluation shows that the consistency between the semantic representations and the learned realizations is high and that the naturalness of the realizations is higher than a hand-crafted baseline.

1 Introduction

One obstacle to the widespread deployment of spoken dialogue systems is the cost involved with hand-crafting the spoken language generation module. Spoken language generation requires a dictionary of mappings between semantic representations of concepts the system wants to express and realizations of those concepts. Dictionary creation is a costly process: an automatic method for creating them would make dialogue technology more scalable. A secondary benefit is that a learned dictionary may produce more natural and colloquial utterances.

We propose a novel method for mining user reviews to automatically acquire a domain specific generation dictionary for information presentation in a dialogue system. Our hypothesis is that reviews that provide individual ratings for various distinguished attributes of review entities can be used to map review sentences to a semantic representation. 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.'*

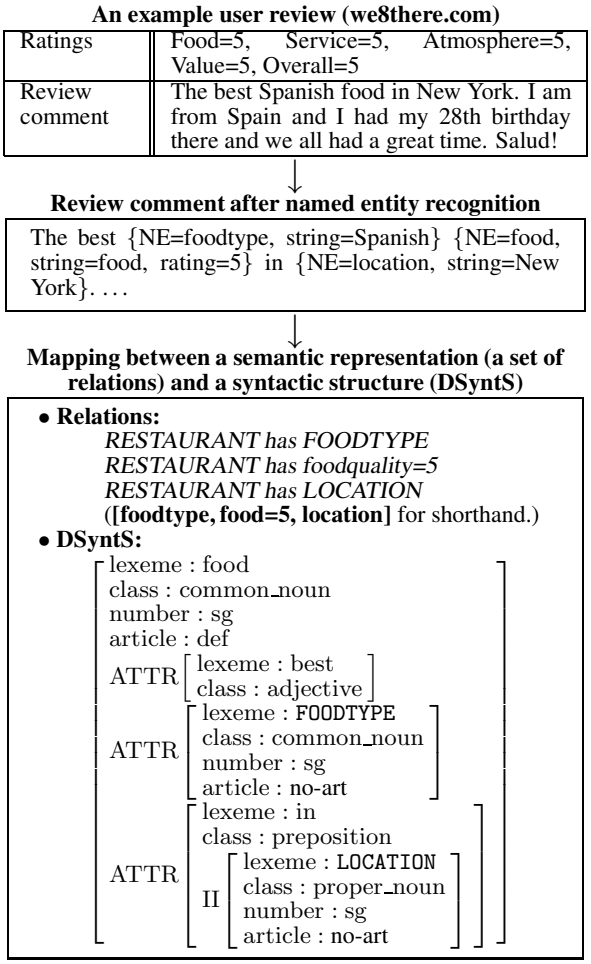

Figure 1: Example of procedure for acquiring a generation dictionary mapping.

We apply the method to extract 451 mappings from restaurant reviews. Experimental analyses show that the mappings learned cover most of the domain ontology, and provide good linguistic variation. A subjective user evaluation indicates that the consistency between the semantic representations and the learned realizations is high and that the naturalness of the realizations is significantly higher than a hand-crafted baseline.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 265–272,
Sydney, July 2006.

Section 2 provides a step-by-step description of the method. Sections 3 and 4 present the evaluation results. Section 5 covers related work. Section 6 summarizes and discusses future work.

2 Learning a Generation Dictionary

Our 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 $\mathcal{S}(\mathcal{U})$. Although templates are widely used in many practical systems (Seneff and Polifroni, 2000; Theune, 2003), we derive syntactic structures to represent the potential realizations, in order to allow aggregation, and other syntactic transformations of utterances, as well as context specific prosody assignment (Walker et al., 2003; Moore et al., 2004).

The method is outlined briefly in Fig. 1 and described below. It comprises 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 webpages' 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.

2.1 Creating the corpus

We created a corpus of restaurant reviews by scraping 3,004 user reviews of 1,810 restaurants posted at we8there.com (http://www.we8there.com/), where each individual review includes a 1-to-5 Likert-scale rating of different restaurant attributes. The corpus consists of 18,466 sentences.

2.2 Deriving semantic representations

The distinguished attributes are extracted from the webpages for each restaurant entity. They include attributes that the users are asked to rate, i.e. *food, service, atmosphere, value*, and *overall*, which have scalar values. In addition, other attributes are extracted from the webpage, such as the *name*, *foodtype* and *location* of the restaurant, which have categorical values. The *name* attribute is assumed to correspond to the restaurant entity. 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 (RESTAURANT) and the distinguished attributes. Thus, the domain ontology consists of the relations:

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

Table 1: Lexicalizations for distinguished attributes.

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

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 problem then is to identify these utterances. We test the hypothesis that, if an utterance $\mathcal{U}$ contains named-entities corresponding to the distinguished attributes, that $\mathcal{R}$ for that utterance includes the relation concerning that attribute in the domain ontology.

We define named-entities for lexicalizations of the distinguished attributes, starting with the seed word for that attribute on the webpage (Table 1).[1] For named-entity recognition, we use GATE (Cunningham et al., 2002), augmented with named-entity lists for locations, food types, restaurant names, and food subtypes (e.g. *pizza*), scraped from the we8there webpages.

We also hypothesize that the rating given for the distinguished attribute specifies the scalar value of the relation. For example, a sentence containing *food* or *meal* 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, e.g. *'*RESTAURANT *has foodquality = 5'* in Fig. 1. Similarly, the other relations in Fig. 1 are assumed to be realized by the utterance "The best Spanish food in New York" because it contains

[1] In future, we will investigate other techniques for bootstrapping these lexicalizations from the seed word on the webpage.

filter	filtered	retained
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

Table 2: Filtering statistics: the number of sentences filtered and retained by each filter.

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 Fig. 1.

2.3 Parsing and DSyntS conversion

We adopt Deep Syntactic Structures (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), and then convert Minipar's representation into DSyntS. Since user reviews are different from the newspaper articles on which Minipar was trained, the output of Minipar can be inaccurate, leading to failure in conversion. We check whether conversion is successful in the filtering stage.

2.4 Filtering

The goal of filtering is to identify $\mathcal{U}$ that realize 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, that $\mathcal{R}$ 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 words indicating that the meaning of $\mathcal{U}$ depends on the surrounding context, that $\mathcal{R}$ will not completely characterizes 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 below eliminate $\mathcal{U}$ that (1) realize semantic relations not in the ontology; (2) contain words indicating that its meaning depends on the context; (3) contain unknown words; or (4) cannot be parsed accurately.

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

Other Relations Filter: The sentence contains named-entities for food subtypes, person names, country names, dates (e.g., today, tomorrow, Aug. 26th) or prices (e.g., 12 dollars), or POS tag CD for numerals. These indicate relations not in the 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), as well as discourse connectives that signal semantic and pragmatic relations of the sentence with other parts of the 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] indicating that its meaning may not be estimated solely by 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.

We apply the filters, in a cascading manner, to the 18,466 sentences with semantic representations. As a result, we obtain 512 (2.8%) mappings of $(\mathcal{U}, \mathcal{R}, \mathcal{S})$. After removing 61 duplicates, 451 distinct (2.4%) mappings remain. Table 2 shows the number of sentences eliminated by each filter.

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

Table 3: Domain coverage of single scalar-valued relation mappings.

3 Objective Evaluation

We evaluate the learned expressions with respect to domain coverage, linguistic variation and generativity.

[2] We used 20 as a threshold.

#	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 4: Counts for multi-relation mappings.

3.1 Domain Coverage

To be usable for a dialogue system, the mappings must have good domain coverage. Table 3 shows the distribution of the 327 mappings realizing a single scalar-valued relation, categorized by the associated rating score.[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.

There are also mappings that express several relations. Table 4 shows the counts of mappings for multi-relation 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 two or three relations, perhaps suggesting that system utterances should not express too many relations in a single sentence.

3.2 Linguistic Variation

We also wish to assess whether the linguistic variation of the learned mappings was greater than what we could easily have generated with a hand-crafted dictionary, or a hand-crafted dictionary augmented with aggregation operators, as in (Walker et al., 2003). Thus, we first categorized the mappings by the patterns of the DSyntSs. Table 5 shows the most common syntactic patterns (more than 10 occurrences), indicating that 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. Furthermore, up to 55% of the learned mappings could be generated from these basic patterns by the application of a combination operator that coordinates multiple adjectives, or coordinates predications over distinct attributes. However, there are 137 syntactic patterns in all, 97 with unique syntactic structures and 21 with two occurrences, accounting for 45% of the learned mappings. Table 6 shows examples of learned mappings with distinct syntactic structures. It would be surprising to see this type of variety in a hand-crafted 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, indicating that the method extracts a wide variety of expressions of varying lengths.

Another interesting aspect of the learned mappings is the wide variety of adjectival phrases (APs) in the common patterns. Tables 7 and 8 show the APs in single scalar-valued relation mappings for *food* and *service* categorized by the associated ratings. Tables for *atmosphere, value* and *overall* can be found in the Appendix. Moreover, the meanings for some of the learned APs are very specific to the particular attribute, e.g. *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 adjectival phrases (APs) in the common patterns 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)).

3.3 Generativity

Our motivation for deriving syntactic representations for the learned expressions was 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 (Walker et al., 2003). We found that only 306 combinations out of a potential 81,318

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

#	syntactic pattern	example utterance	count	ratio	accum.
1	NN VB JJ	The atmosphere is wonderful.	92	20.4%	20.4%
2	NN VB RB JJ	The atmosphere was very nice.	52	11.5%	31.9%
3	JJ NN	Bad service.	36	8.0%	39.9%
4	NN VB JJ CC JJ	The food was flavorful but cold.	25	5.5%	45.5%
5	RB JJ NN	Very trendy ambience.	22	4.9%	50.3%
6	NN VB JJ CC NN VB JJ	The food is excellent and the atmosphere is great.	13	2.9%	53.2%
7	NN CC NN VB JJ	The food and service were fantastic.	10	2.2%	55.4%

Table 5: Common syntactic patterns of DSyntSs, flattened to a POS sequence for readability. NN, VB, JJ, RB, CC stand for noun, verb, adjective, adverb, and conjunction, respectively.

[overall=1, value=2] Very disappointing experience for the money charged.
[food=5, value=5] The food is excellent and plentiful at a reasonable price.
[food=5, service=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] An absolutely outstanding value with fantastic FOODTYPE food.
[food=5, foodtype, location, overall=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=4] Very pleasantly surprised by the food.
[food=1] The food has gone downhill.
[atmosphere=5, overall=5] This is a quiet little place with great atmosphere.
[atmosphere=5, food=5, overall=5, service=5, value=5] The food, service and ambience of the place are all fabulous and the prices are downright cheap.

Table 6: Acquired generation patterns (with shorthand for relations in square brackets) whose syntactic patterns occurred only once.

combinations (0.37%) were successful. This is because the merge operation in SPaRKy requires that the subjects and the verbs of the two DSyntSs are identical, e.g. the subject is RESTAURANT and verb is *has*, whereas the learned DSyntSs often place the attribute in subject position as a definite noun phrase. However, the learned DSyntS can be incorporated into SPaRKy using the semantic representations to substitute learned DSyntSs into nodes in the sentence plan tree. Figure 2 shows some example utterances generated by SPaRKy with its original dictionary and example utterances when the learned mappings are incorporated. The resulting utterances seem more natural and colloquial; we examine whether this is true in the next section.

4 Subjective Evaluation

We evaluate the obtained mappings in two respects: the consistency between the automatically derived semantic representation and the realization, and the naturalness of the realization.

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

Table 7: Adjectival phrases (APs) in single scalar-valued relation mappings for *foodquality*.

For comparison, we used a baseline of handcrafted mappings from (Walker et al., 2003) except that we changed the word *decor* to *atmosphere* and added 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.*"

The learned mappings include 23 distinct semantic representations for a single-relation (22 for scalar-valued relations and one for location) and 50 for multi-relations. Therefore, using the handcrafted mappings, we first created 23 utterances for the single-relations. We then created three utterances for each of 50 multi-relations using different clause-combining operations from (Walker et al., 2003). This gave a total of 173 baseline utterances, which together with 451 learned mappings,

service=1	awful, bad, great, horrendous, horrible, inattentive, forgetful and slow, marginal, really slow, silly and inattentive, still marginal, terrible, young
service=2	overly slow, very slow and inattentive
service=3	bad, bland and mediocre, friendly and knowledgeable, good, pleasant, prompt, very friendly
service=4	all very warm and welcoming, attentive, extremely friendly and good, extremely pleasant, fantastic, friendly, friendly and helpful, good, great, great and courteous, prompt and friendly, really friendly, so nice, swift and friendly, very friendly, very friendly and accommodating
service=5	all courteous, excellent, excellent and friendly, extremely friendly, fabulous, fantastic, friendly, friendly and helpful, friendly and very attentive, good, great, great, prompt and courteous, happy and friendly, impeccable, intrusive, legendary, outstanding, pleasant, polite, attentive and prompt, prompt and courteous, prompt and pleasant, quick and cheerful, stupendous, superb, the most attentive, unbelievable, very attentive, very congenial, very courteous, very friendly, very friendly and helpful, very friendly and pleasant, very friendly and totally personal, very friendly and welcoming, very good, very helpful, very timely, warm and friendly, wonderful

Table 8: Adjectival phrases (APs) in single scalar-valued relation mappings for *servicequality*.

yielded 624 utterances for evaluation.

Ten subjects, all native English speakers, evaluated the mappings by reading them from a webpage. For each system utterance, 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.

4.1 Results

Table 9 shows the means and standard deviations of the scores for baseline vs. learned utterances for consistency and naturalness. A t-test shows that the consistency of the learned expression is significantly lower than the baseline (df=4712, $p < .001$) but that their naturalness is significantly higher than the baseline (df=3107, $p < .001$). However, consistency is still high. Only 14 of the learned utterances (shown in Tab. 10) have a mean consistency score lower than 3, which indicates that, by and large, the human judges felt that the inferred semantic representations were consistent with the meaning of the learned expressions. The correlation coefficient between consistency and naturalness scores is 0.42, which indicates that consistency does not greatly relate to naturalness.

Original SPaRKy utterances

- Babbo has the best overall quality among the selected restaurants with excellent decor, excellent service and superb food quality.
- Babbo has excellent decor and superb food quality with excellent service. It has the best overall quality among the selected restaurants.

↓

Combination of SPaRKy and learned DSyntS

- Because **the food is excellent, the wait staff is professional and the decor is beautiful and very comfortable**, Babbo has the best overall quality among the selected restaurants.
- Babbo has the best overall quality among the selected restaurants because **atmosphere is exceptionally nice, food is excellent and the service is superb**.
- 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.

Figure 2: Utterances incorporating learned DSyntSs (Bold font) in SPaRKy.

	baseline		learned		stat.
	mean	sd.	mean	sd.	sig.
Consistency	**4.714**	0.588	**4.459**	0.890	+
Naturalness	**4.227**	0.852	**4.613**	0.844	+

Table 9: Consistency and naturalness scores averaged over 10 subjects.

We also performed an ANOVA (ANalysis Of VAriance) of the effect of each relation in $\mathcal{R}$ on naturalness and consistency. There were no significant effects except that mappings combining food, service, and atmosphere were significantly worse (df=1, F=7.79, p=0.005). However, there is a trend 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, suggesting that perhaps it is easier to learn some mappings than others.

5 Related Work

Automatically finding sentences with the same meaning has been extensively studied in the field of automatic paraphrasing using parallel corpora and corpora with multiple descriptions of the same events (Barzilay and McKeown, 2001; Barzilay and Lee, 2003). Other work finds predicates of similar meanings by using the similarity of contexts around the predicates (Lin and Pantel, 2001). However, these studies find a set of sentences with the same meaning, but do not associate a specific meaning with the sentences. One exception is (Barzilay and Lee, 2002), which derives mappings between semantic representations and realizations using a parallel (but unaligned) corpus consisting of both complex semantic input and corresponding natural language verbalizations for mathemat-

shorthand for relations and utterance	score
[food=4] The food is delicious and beautifully prepared.	2.9
[overall=4] A wonderful experience.	2.9
[service=3] The service is bland and mediocre.	2.8
[atmosphere=2] The atmosphere here is eclectic.	2.6
[overall=3] Really fancy place.	2.6
[food=3, service=4] Wonderful service and great food.	2.5
[service=4] The service is fantastic.	2.5
[overall=2] The RESTAURANTNAME is once a great place to go and socialize.	2.2
[atmosphere=2] The atmosphere is unique and pleasant.	2.0
[food=5, foodtype] FOODTYPE and FOODTYPE food.	1.8
[service=3] Waitstaff is friendly and knowledgeable.	1.7
[atmosphere=5, food=5, service=5] The atmosphere, food and service.	1.6
[overall=3] Overall, a great experience.	1.4
[service=1] The waiter is great.	1.4

Table 10: The 14 utterances with consistency scores below 3.

ical proofs. However, our technique does not require parallel corpora or previously existing semantic transcripts or labeling, and user reviews are widely available in many different domains (See *http://www.epinions.com/*).

There is also significant previous work on mining user reviews. For example, Hu and Liu (2005) use reviews to find adjectives to describe products, and Popescu and Etzioni (2005) automatically find features of a product together with the polarity of adjectives used to describe them. They both aim at summarizing reviews so that users can make decisions easily. Our method is also capable of finding polarities of modifying expressions including adjectives, but on a more fine-grained scale of 1 to 5. However, it might be possible to use their approach to create rating information for raw review texts as in (Pang and Lee, 2005), so that we can create mappings from reviews without ratings.

6 Summary and Future Work

We proposed automatically obtaining mappings between semantic representations and realizations from reviews with individual ratings. The 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; and (3) the naturalness of the realizations are significantly higher than the baseline.

There are also limitations in our method. Even though consistency is rated highly by human subjects, this may actually be a judgement 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 7, cannot be guaranteed to be synonymous, which may be required for use in spoken language generation. Rather, an examination of the adjectival phrases in Table 7 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 the flavors. This suggests that perhaps the ontology would benefit from representing these sub-attributes of the food attribute, and sub-attributes in general. Another problem with consistency is that the same AP, e.g. *very good* in Table 7 may appear with multiple ratings. For example, *very good* is used for every foodquality rating from 2 to 5. Thus some further automatic or by-hand analysis is required to refine what is learned before actual use in spoken language generation. Still, our method could reduce the amount of time a system designer spends developing the spoken language generator, and 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 2.4% 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 here has high precision, and automatic techniques may introduce noise. A related 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 sentences 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 webpages may not cover all of 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 possibly help here, although these techniques currently have the limitation that they 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 these would have to be constructed by hand, as in current practice, or perhaps other types of corpora or resources could be utilized. In addition, 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). Finally, while we believe that this technique will apply across domains, it would be useful to test it on domains such as movie reviews or product reviews, which have more complex domain ontologies.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was supported by a Royal Society Wolfson award to Marilyn Walker and a research collaboration grant from NTT to the Cognitive Systems Group at the University of Sheffield.

References

Regina Barzilay and Lillian Lee. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proc. EMNLP*, pages 164–171.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proc. HLT/NAACL*, pages 16–23.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. 39th ACL*, pages 50–57.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. 40th ACL*.

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

Julia Hirschberg and Diane. J. Litman. 1987. Now let's talk about NOW: Identifying cue phrases intonationally. In *Proc. 25th ACL*, pages 163–171.

Minqing Hu and Bing Liu. 2005. Mining and summarizing customer reviews. In *Proc. KDD*, pages 168–177.

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

Benoit Lavoie and Owen Rambow. 1997. A fast and portable realizer for text generation systems. In *Proc. 5th Applied NLP*, pages 265–268.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Dekang Lin. 1998. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*.

Johanna D. Moore, Mary Ellen Foster, Oliver Lemon, and Michael White. 2004. Generating tailored, comparative descriptions in spoken dialogue. In *Proc. 7th FLAIR*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. 43st ACL*, pages 115–124.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. HLT/EMNLP*, pages 339–346.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proc. Corpus Linguistics Workshop on Using Corpora for NLG*.

Stephanie Seneff and Joseph Polifroni. 2000. Formal and natural language generation in the mercury conversational system. In *Proc. ICSLP*, volume 2, pages 767–770.

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

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th ACL*, pages 417–424.

Marilyn Walker, Rashmi Prasad, and Amanda Stent. 2003. A trainable generator for recommendations in multimodal dialog. In *Proc. Eurospeech*, pages 1697–1700.

Michael White and Ted Caldwell. 1998. EXEMPLARS: A practical, extensible framework for dynamic text generation. In *Proc. INLG*, pages 266–275.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proc. AAAI*, pages 761–769.

Appendix

Adjectival phrases (APs) in single scalar-valued relation mappings for *atmosphere, value*, and *overall*.

atmosphere=2	eclectic, unique and pleasant
atmosphere=3	busy, pleasant but extremely hot
atmosphere=4	fantastic, great, quite nice and simple, typical, very casual, very trendy, wonderful
atmosphere=5	beautiful, comfortable, excellent, great, interior, lovely, mellow, nice, nice and comfortable, phenomenal, pleasant, quite pleasant, unbelievably beautiful, very comfortable, very cozy, very friendly, very intimate, very nice, very nice and relaxing, very pleasant, very relaxing, warm and contemporary, warm and very comfortable, wonderful
value=3	very reasonable
value=4	great, pretty good, reasonable, very good
value=5	best, extremely reasonable, good, great, reasonable, totally reasonable, very good, very reasonable
overall=1	just bad, nice, thoroughly humiliating
overall=2	great, really bad
overall=3	bad, decent, great, interesting, really fancy
overall=4	excellent, good, great, just great, never busy, not very busy, outstanding, recommended, wonderful
overall=5	amazing, awesome, capacious, delightful, extremely pleasant, fantastic, good, great, local, marvelous, neat, new, overall, overwhelmingly pleasant, pampering, peaceful but idyllic, really cool, really great, really neat, really nice, special, tasty, truly great, ultimate, unique and enjoyable, very enjoyable, very excellent, very good, very nice, very wonderful, warm and friendly, wonderful