Inferring the Environment in a Text-to-Scene Conversion System

Richard Sproat
Human/Computer Interaction Research
AT&T Labs — Research
180 Park Avenue
Florham Park, NJ 07932, USA
rws@research.att.com

Abstract

There has been a great deal of work over the past decade on inferring semantic information from text corpora. This paper is another instance of this kind of work, but is also slightly different in that we are interested not in extracting semantic information per se, but rather real-world knowledge. In particular, given a description of a particular action — e.g. *John was eating breakfast* — we want to know where John is likely to be, what time of day it is, and so forth. Humans on hearing this sentence would form a mental image that makes a lot of inferences about the *environment* in which this action occurs: they would probably imagine someone in their kitchen in the morning, perhaps in their dining room, seated at a table, eating a meal.

We propose a method that makes use of Dunning's likelihood ratios to extract from text corpora strong associations between particular actions and locations or times when those actions occur. We also present an evaluation of the method. The context of this work is a text-to-scene conversion system called WordsEye, where in order to depict an action such as *John was eating breakfast*, it is desirable to make reasonable inferences about where and when that action is taking place so that the resulting picture is a reasonable match to one's mental image of the action.

Keywords

Common sense knowledge; statistical natural language processing; text-to-scene conversion.

INTRODUCTION

There has been a great deal of work over the past decade on inferring semantic information from text corpora; see [9, 8, 17, 18, 2, 19, 12, 13] for some examples. This paper is another instance of this kind of work, but is also slightly different in that we are interested not in extracting semantic information per se, but rather real-world knowledge. In particular, given a description of a particular action — e.g. *John was eating breakfast* — we want to know where John is likely to be, what time of day it is, and so forth. Humans on hearing this sentence would probably form a mental image of someone in their kitchen, perhaps in their dining room, seated at a table, eating a meal in the morning. But note that the sentence omits a lot of this information, and says nothing explicit about the location of the action, or the time of day. Nonetheless, people would usually make these inferences about the *environment* in which the particular action occurs.

The context of this work is a text-to-scene conversion system called WordsEye, which we describe in the next section. Subsequent sections describe the method for extracting information about the environment from text corpora, and an evaluation of the method.

K-CAP'01, October 22-23, 2001, Victoria, British Columbia, Canada.

THE WORDSEYE SYSTEM

WordsEye [5] is a system for converting from English text into three-dimensional graphical scenes that represent that text. WordsEye works by performing syntactic and semantic analysis on the input text, producing a description of the arrangement of objects in a scene. An image is then generated from this scene description. At the core of WordsEye is the notion of a "pose", which can be loosely defined as a figure (e.g. a human figure) in a configuration suggestive of a particular action. For example a human figure holding an object in its hand in a throwing position would be a pose that suggests actions such as *throw* or *toss*. Substituting for the figure or the object will allow one to depict different statements, such as *John threw the egg* or *Mary tossed the small toy car*.

The natural language component in the current incarnation of WordsEye is built in part on several already existing components, including Church's [3] part of speech tagger, Collins' head-driven stochastic parser [4] and the WordNet semantic hierarchy [7]. The parsed sentence is first converted into a dependency representation. Then lexical semantic rules are applied to this dependency representation to derive the com-

Figure 1: *Mary uses the crossbow. She rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces Mary. A gigantic teacup is in front of the store. The gigantic mushroom is in the teacup. The castle is to the right of the store.*

Figure 2: *John was eating breakfast.* **The light in the sky coming through the window is morning light (though that may be hard to see in a black and white version).**

ponents of the scene description. For instance the verb *throw* invokes a semantic rule that constructs a scene component representing an action (ultimately mapped to a pose) where the lefthand noun phrase dependent represents an actor, the righthand noun phrase dependent a patient, and some dependent prepositional phrases the path of the patient.[1] WordNet is used as part of the implementation of noun semantics, both to derive appropriate sets of objects (e.g. *the vehicle* will get all vehicle objects by inheritance from WordNet subclasses such as *car*, *airplane*, etc.); and in subsequent reference resolution (so that one can refer to, e.g., an *allosaurus* and subsequently use *dinosaur* to refer to the previously evoked allosaurus).

The depiction module of WordsEye interprets the scene description to produce a set of low-level depictors representing poses, spatial relations, color attributes etc. Transduction rules are applied to resolve conflicts and add implict constraints. The resulting depictors are then used (while maintaining constraints) to manipulate the 3D objects that constitute the final, renderable 3D scene. An example of a fairly complex scene constructed with WordsEye is shown in Figure 1.

One problem that arises in such a system is how to derive the large amount of knowledge that is needed in order to give reasonable depictions. Suppose I say: *John was driving to the store.* In understanding this sentence and visualizing what it means, a human would probably assume that John was in the driver's seat of a car, on a road, possibly passing buildings, and so forth. Many of these inferences are defeasible: I can easily cancel the inference about the road, for example, by saying *John was driving to the store across the muddy field.* But without such explicit cancellation the inferences seem fairly robust. To take another example, if we say *John ate his dinner at 7*, we assume that it is 7 in the evening (possibly near twilight), that he is in a room such as his dining room or his kitchen (or possibly in a restaurant), and that he is seated at a table. Or if *John was eating breakfast*, we would usually assume that it is morning, and that John is in his kitchen or dining room. See Figure 2. Finally, if *John is shoveling snow*, it is probably winter.

Some of this knowledge is represented in WordsEye as part of the word's meaning. For example, the depiction phase of WordsEye knows that for *drive*, the driver should be using some sort of vehicle, and will select an appropriate vehicle and place the driver in the driver's seat. But other common sense knowledge is more tenuously linked: if John is washing his face, he is probably in a bathroom, but need not be: there is nothing in the meaning of *wash face*, that implies a bathroom.

An important problem is how to acquire this kind of knowledge. One approach would of course be to do it by hand, possibly making use of already hand-built ontologies such as Cyc [14], or Mikrokosmos [15]. In this paper we explore the alternative of deriving this kind of information from text corpora.[2]

The question posed by this paper can therefore be stated as follows: if John is eating dinner, can we infer from text corpora where he is and what time of day it is? If John is raking leaves, can we infer from text corpora what season it is and where he is likely to be?

METHOD

The first step involves computing a set of concordance lines for terms that can denote elements of the set of interest. For

[1] We have just started investigating the use of FrameNet [10] for verbal semantics.

[2] We do not mean to imply, however, that hand-built ontologies such as Cyc and statistical methods such as the one proposed here, are at odds with one another. Rather, the two approaches complement one another, as we will suggest in the final section.

example, if one is interested in activities that can take place in various rooms of a house, one would compute concordance lines for terms like *kitchen*, *living room*, *dining room*, *hallway*, *laundry room* and so forth: so, for the key word *kitchen*, one would simply find all places in a corpus that have the word *kitchen*, and for each of these, output a line containing the key word surrounded by words in a predetermined window of that corpus location.

We used a corpus of 415 million words of English text, consisting of about nine years of the *Associated Press* newswire, the Bible, the Brown corpus [11], *Grolier's Encyclopedia*, about 70 texts of various kinds published by Harper and Row, about 2.7 million words of psychiatry texts, a corpus of short movie synopses, and 62 million words of the *Wall Street Journal*. The texts in this corpus had already been automatically tagged with a part of speech tagger [3] and so the concordance lines also contain part of speech information for each word.[3]

Sample concordance lines for various rooms from the 1996 *Associated Press* newswire are given in Figure 3. (Here we omit the part of speech information for readability.) As expected, the data are noisy: for example in the third line, *Kitchen* is a family name, not a room in the house. Note that in the actual implementation, a window of 40 words on each side of the target is used, wider than what is shown here.

Once the concordance lines are collected, and after sorting to remove duplicates (newswire text especially contains a lot of repeated stories), we extract *verb-object* (e.g. *wash face*) and *verb-preposition-object* (e.g. *get into bed*) tuples. Unlike verbs alone, verb-argument tuples of this kind are usually pretty good indicators of a particular action. Thus, whereas *wash* is consistent with many activities (e.g. washing oneself, washing one's car, washing clothes), a particular verb-object construction such as *wash clothes* is usually indicative of a particular activity. In the present system, the tuples are extracted using a simple matching algorithm that looks for verbal part-of-speech tags and then searches for what looks like the end of the following noun phrase, with a possible intervening preposition.[4] Verb-object and verb-preposition-object tuples extracted from the concordance lines in Figure 3 are shown in Figure 4.

Once again the data are noisy, and include misanalyses (*didn't window*) and complex nominals that are not instances of verb-object constructions (*swimming pool*). Apart from misanalyses of the tuples, one also finds many instances where the target term does not have the intended denotation. For example, a concordance line matching *kitchen* will not always have to do with kitchens. As we saw above, *Kitchen* may be a family name, but a more common instance is that it is part of a complex nominal, such as *kitchen knife*. In such instances the text is not generally talking about kitchens, but rather about kitchen knives, which can be used in rooms besides kitchens. To remove such cases we filter the concordance lines to remove the most frequent collocations (for example the 200 most frequent ones).

The next step is to compute the association between each of the tuples and the target term, such as the name of a room. For this stage we use likelihood ratios [6, 16], which compute the relative likelihood of two hypotheses concerning two events e_1 and e_2:

- Hypothesis 1: $p(e_2|e_1) = p = p(e_2|\neg e_1)$
- Hypothesis 2: $p(e_2|e_1) = p_1 \neq p_2 = p(e_2|\neg e_1)$

Hypothesis 1 simply says that the probability of e_2 occurring given e_1 is indistinguishable from the probability of e_2 occurring given something other than e_2: i.e., the e_2 is not particularly expected (or unexpected) given e_1. Hypothesis 2 says, in contrast, that there is a difference in expectation, and that e_2 is dependent on e_1.

We can estimate the probabilities p, p_1 and p_2 by the maximum likelihood estimate as follows, where c_1, c_2 and c_{12} are, respectively the frequency of e_1, of e_2, and of e_1 and e_2 cooccurring; and N is the size of the corpus:

$$p = \frac{c_2}{N}, p_1 = \frac{c_{12}}{c_1},$$
$$p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

If we assume a binomial distribution

$$b(k; n, x) = \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{(n-k)}$$

then the likelihoods of the two hypotheses, given the observed counts e_1, e_2 and e_{12}, can be computed as:

$$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)$$
$$L(H_2) = b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2)$$

The *log* likelihood ratio for the two hypotheses then reduces as follows:

$$\begin{aligned} log\lambda &= log\frac{L(H_1)}{L(H_2)} \\ &= logL(c_{12}, c_1, p) + logL(c_2 - c_{12}, N - c_1, p) \\ &\quad -logL(c_{12}, c_1, p_1) - logL(c_2 - c_{12}, N - c_1, p_2) \end{aligned}$$

where:

[3] The concordance itself is computed using a corpus encoding representation and a set of corpus tools developed at AT&T, but this could just as easily have been done with any of a number of other concordancing software packages.

[4] This is currently done with an ad hoc script, though we are investigating using Cass [1], a robust chunk parser, in the future. Note that while the Collins parser is used in the runtime version of WordsEye, it is far too slow to use to parse large amounts of text.

anything else, her books are about memories:	*kitchen*	memories, barnyard memories, family memories
both videotapes and photos of her in bathrooms and	*bedroom*	and asks for an unspecified amount of
will happen to Mr Tarkanian," said Jack	*Kitchen*	, one of the NCAA's lawyers
grounded for telling his parents he didn't open his	*bedroom*	window. He confessed in
gone, replaced by a big house with five	*bathroom*	and an indoor swimming pool.
The second child was born in a	*bedroom*	of their home near Scottsdale after Corvin
beds in semiprivate rooms at one end of a	*hallway*	separated from the "older adult"
and the couple's 15-month-old son use a downstairs	*bedroom*	that lies in Granite City along with
of the halls, equipped with microwaves and other	*kitchen*	appliances not allowed in individual rooms.

Figure 3: Sample concordance lines from the 1996 Associated Press.

asks		amount	**bedroom**
happen	to	Tarkanian	**kitchen**
grounded		parents	**bedroom**
telling		parents	**bedroom**
didn't		window	**bedroom**
replaced	by	house	**bathroom**
swimming		pool	**bathroom**
born	in	home	**bedroom**
use	in	City	**bedroom**
lies	in	City	**bedroom**
equipped	with	microwaves	**kitchen**
allowed	in	rooms	**kitchen**

Figure 4: Some verb-argument combinations extracted from Figure 3.

$$L(k, n, x) = x^k(1 - x)^{n-k}$$

Following [6, 16] we make use of the fact that $-2log\lambda$ is asymptotically χ^2 distributed, and compute $-2log\lambda$, rather than just $log\lambda$. In what follows we assume p value of 0.05, which has a critical χ^2 value of 3.84 for one degree of freedom. Thus any $-2log\lambda$ value of 3.84 or above will be considered evidence of association.

After the likelihood ratios for each tuple-term pair are computed, we then sort the tuple-term pairs, and filter to remove those that are below the significance threshold; in the process of doing this, we also lemmatize the verb forms, or in other words replace inflected verbs (e.g. *eats*) by their base forms (e.g. *eat*). A sample of the highest ranking tuple-term associations is given in Figure 5. Again, there is still noise, including a misanalyzed complex nominal (*dine room* from *dining room*), misparsed examples (*find in Simpson* from *find in Simpson's X*) and so forth.

The final stage is to filter the list for tuples that designate reasonable depictable actions. We do this by extracting activities from the set of sentences input to the WordsEye system; at the time of writing this consisted of 20K words (about 3,400 sentences). We then use these activities to filter the raw likelihood-ratio-ordered list. An example of a filtered list is shown in Figure 6. A similar example for times of day is shown in Figure 7.

EVALUATION

The system has been evaluated by human subjects on its predictions for the **rooms**, **seasons** and **times** of day in which particular actions or situations occur. Clearly these are not the only things that one would like to infer about a scene, but they are three fairly obvious ones, and serve to give us a metric for evaluating the method as a whole.

The test used sentences constructed based on verb-object or verb-preposition-object tuples from the final filtered lists for rooms, seasons and times, as described in the last section. This meant that the system would be able to predict an answer for at least one of these categories for each of the sentences, but at the same time there was no guarantee that the prediction would be correct. This resulted in 106 sentences from which 90 were randomly selected. Of these 90, 30 were submitted to the system to label the choices for the three variables listed above; 30 were given to a human for labeling; and 30 — the "baseline" system — had the answers labeled randomly.

The three sets of judgments were randomized, and presented via a web-based interface to subjects, who were informed that they were judging the output of an automatic system; subjects were not informed that some sentences had been labeled randomly, or that some had been labeled by a human. (For those who are interested, the exact instructions given to subjects are shown in the Appendix.)

306.585215	143	424	10227	dine room	dining room
196.753628	63	65	32243	find in Simpson	bedroom
150.457758	29	31	10227	serve in room	dining room
137.680378	35	51	10227	designate areas	dining room
117.189848	23	25	10227	eat in room	dining room
109.719646	25	29	12457	wash clothes	laundry room
107.275571	24	30	10227	cook on premises	dining room
100.616896	19	19	12457	sell in America	laundry room
96.602198	205	575	32243	live room	bedroom
79.429912	15	15	12457	cook appliances	laundry room
76.659647	43	68	28224	kill people	garage
61.528933	49	64	51214	sit at table	kitchen
61.103395	30	47	24842	give birth	bathroom
61.067298	18	18	32243	see socks	bedroom
58.542468	16	16	28224	rent van	garage
54.146381	18	21	24842	wash hands	bathroom
51.280771	21	54	10227	dine rooms	dining room
51.111709	26	28	51214	prepare meals	kitchen
49.807875	10	10	14575	push down gantlet	hallway
49.807875	10	10	14575	form gantlet	hallway
47.564595	13	13	28224	carry bomb	garage

Figure 5: Most likely actions associated with particular rooms. Columns represent, from left to right: the likelihood ratio; the frequency of the tuple/target-term pair; the frequency of the tuple; the frequency of the target term; the tuple; and the target term.

The full set of choices for each of the categories were as follows:

Room: bedroom, kitchen, laundry room, living room, bathroom, hallway, garage, ANY, NONE OF ABOVE

Time of day: morning, midday, afternoon, evening, night, ANY

Season: winter, spring, summer, autumn, ANY

"ANY" indicates that any of the choices would be acceptable. "NONE OF ABOVE", in the case of rooms, indicates that the action could not occur in any of the rooms; typically this would be because the action occurs outside.

The interpretation of the subjects' judgments are as follows:

- If the preselected choice is left alone it is assumed *correct*.
- If the preselected choice is changed to "ANY", it is assumed that the preselected choice *may be* okay.
- If the preselected choice is changed to any other selection, it is assumed to be *incorrect*.

63 subjects, all employees at AT&T Labs, participated in the experiment. Subjects were rewarded with a bar of chocolate of their choice. Results are presented in Table 1. In this table, "human" denotes the 30 human-judged sentences; "system" the sentences tagged by the system; and "baseline" the randomly tagged 30 sentences. Note that since we are dealing with three predictions in each case (room, season and time of day), we have a total of 90 judgments for each of the human, system and baseline conditions. For each condition we report total errors, and *real errors*, which are errors where the subject changed the setting to something other than "ANY". As indicated in the table, all differences between the baseline and the system were significant at at least the $p < 0.01$ level on a two-sample t-test, except for the real errors for times, which were significant at the $p < 0.05$ level. All differences between the human and the system were significant at at least the $p < 0.01$ level.

So while the system does significantly better than the random baseline, it also does significantly worse than a human, and there is thus still room for improvement. An interesting general point is that the strength of the judgments seem to vary greatly across the three types — rooms, times and seasons. The subjects marked the most errors for rooms, but fewer errors for time of day or season. This suggests that, at least for the kinds of actions studied here, getting the location right is more important than getting the time of day or season right.

CONCLUSIONS & FUTURE WORK

The method we describe in this paper is fully implemented and forms part of the WordsEye system. While the technique clearly works better than a random assignment of environmental properties, it is still significantly worse than human judgments, so a major component of future work will be trying to close this gap. We are also working to expand the cov-

92.282095	175	433	24730	live room	bedroom
73.256801	17	21	7906	wash clothes	laundry room
51.118373	18	20	21056	wash hands	bathroom
35.438165	19	26	23479	drive car	garage
34.289413	18	26	21056	go to bathroom	bathroom
30.699638	16	23	21056	brush teeth	bathroom
16.510044	5	5	23479	run car	garage
16.107447	18	29	32408	wash dishes	kitchen
14.545979	4	6	7906	go to store	laundry room
14.284725	11	18	24730	go to bed	bedroom
13.490176	10	18	21056	take shower	bathroom
13.286761	5	5	32408	see in kitchen	kitchen
12.792577	4	4	24730	sit on sofa	bedroom
11.718897	11	20	24730	sit on bed	bedroom
10.559389	3	3	21056	sit on toilet	bathroom
10.329526	9	13	32408	sit at table	kitchen
9.594336	3	3	24730	hold knife	bedroom
9.594336	3	3	24730	climb over wall	bedroom
8.774370	5	11	12756	sit on floor	hallway
8.495289	5	6	32408	make breakfast	kitchen
8.240026	4	5	24730	play guitar	bedroom
8.177386	6	8	32408	eat meal	kitchen
7.971921	3	3	32408	cook meal	kitchen
7.945854	11	24	24730	leave house	bedroom
7.945854	11	24	24730	knock on door	bedroom

Figure 6: Most likely actions associated with particular rooms, after filtering with tuples extracted from the WordsEye input sentences. Columns are as in Figure 5.

erage of the technique, in particular by investigating other (implicit) features of the environment that can be predicted by corpus-based methods.

The evaluation of the system reported here evaluates only descriptions for which the system can make a prediction: in effect, then, we have considered the *precision* of the method. What we do not have a measure for at present is the *recall*, or in other words the percentage of sentences for which one ought to make a prediction, but for which we do not have the data to do so. In future work we hope to be able to say more about recall, and propose measures for evaluating it.

Finally, the work reported here considers only classes of information — time of day, location, and so forth — which were selected by hand. As reviewers have noted, and as we are aware, this is not ideal: one would like to be able to let a method loose on a large text corpus, and have it learn interesting associations of all kinds, not just associations that we happened to think of. At present it is not clear how to do this. One thing that is clear is that the more unconstrained the search is, the more "sanity checks" will have to be in place to make sure that the system does not learn useless or absurd associations. It is possible, as some reviewers have noted, that large ontologies such as Cyc may be of use in filtering more absurd associations.

ACKNOWLEDGMENTS

I thank Owen Rambow, Bob Coyne and Candy Kamm for suggestions and help with setting up the evaluation experiment. I also thank Fritz Lehmann, Doug Lenat, and two anonymous reviewers for useful comments.

REFERENCES

1. S. Abney. Partial parsing via finite-state cascades. In J. Carroll, editor, *Workshop on Robust Parsing*, pages 8–15. ESSLLI, Prague, 1996.

2. M. Berland and E. Charniak. Finding parts in very large corpora. In *Proceedings of the North American ACL*, pages 57–64, College Park, MD, 1999.

3. K. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143. Association for Computational Linguistics, 1988.

4. M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, Philadelphia, PA, 1999.

5. B. Coyne and R. Sproat. WordsEye: An automatic text-

35.729439	28	40	385312	read newspapers	morning
33.804553	32	50	385312	eat breakfast	morning
26.415083	15	32	166691	drink tea	evening
19.529204	38	48	743374	sleep on floor	night
17.679023	13	18	385312	look in mirror	morning
13.972083	7	8	385312	celebrate Easter	morning
11.686620	8	8	743374	play trumpet	night
11.240171	10	28	176501	eat lunch	afternoon
10.322243	126	213	743374	go to bed	night
9.572043	15	60	166691	eat dinner	evening
9.413257	6	14	166691	cook meal	evening
9.232992	16	32	385312	take shower	morning
8.673155	2	2	176501	see boat	afternoon
8.673155	2	2	176501	roll in front	afternoon
8.673155	2	2	176501	rake leaves	afternoon
8.573500	2	3	71317	sleep in chair	noon
8.325358	3	3	385312	throw egg	morning
8.325358	3	3	385312	take to hills	morning
7.824066	17	22	743374	sleep in bed	night

Figure 7: Most likely actions associated with particular times of day, after filtering with tuples extracted from the WordsEye input sentences. Columns are as in Figure 5.

	ERR	Real ERR	rERR	Real rERR	sERR	Real sERR	tERR	Real tERR
H	8.2** (6.2)	3.7** (3.3)	4.2** (3.0)	2.7** (2.4)	1.0** (1.1)	0.1** (0.3)	3.2** (3.0)	1.1** (1.7)
S	22.8 (8.3)	12.3 (5.7)	10.9 (2.7)	8.0 (2.4)	6.6 (2.6)	1.5 (0.9)	5.7 (3.8)	3.2 (3.7)
B	56.4** (19.7)	27.4** (7.9)	23.0** (4.4)	21.0** (3.8)	18.1** (6.8)	2.3** (1.2)	16.4** (8.7)	4.6* (4.2)

Table 1: Results of an evaluation on rooms, seasons, and times of day. "H(uman)" denotes the 30 human-judged sentence; "S(ystem)" the sentences tagged by the system; and "B(aseline)" the randomly tagged 30 sentences. Shown are the means and (standard deviations) for 63 subjects for: total errors ERR (= all changes including "ANY"); total real errors (= changes not counting "ANY"); and total and real errors for rooms (rERR), seasons (sERR) and times (tERR). Differences between the baseline and the system were significant at at least the $p < 0.01$ level on a two-sample t-test where indicated with a '' on the mean for the baseline, and at the $p < 0.05$ level otherwise. Significances for the difference between the system and human are similarly indicated with asterisks on the means for the human-assigned judgments.**

to-scene conversion system. In *SIGGRAPH 2001*, Los Angeles, CA, 2001.

6. T. E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

7. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

8. M. Hearst. Automatic acquisition of hyponyms for large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING)*, Nantes, France, 1992.

9. D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA, 1990. ACL.

10. C. Johnson, C. Fillmore, E. Wood, J. Ruppenhofer, M. Urban, M. Petruck, and C. Baker. The FrameNet project: Tools for lexicon building, version 0.7. Technical report, International Computer Science Institute, University of California, Berkeley, Berkeley, CA, 2001. `www.icsi.berkeley.edu/~framenet/book.html`.

11. H. Kucera and W. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, 1967.

12. M. Lapata. The automatic interpretation of nominalizations. In *Proceedings ot the 17th National Conference on Artificial Intelligence (AAAI)*, pages 716–721, Austin, TX, 2000.

13. M. Lapata. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Pro-*

ceedings of the North American ACL, Pittsburgh, PA, 2001.

14. D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), November 1995.

15. K. Mahesh and S. Nirenburg. Semantic classification for practical natural language processing. In *Proceedings of the Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting*, Chicago, IL, October 1995.

16. C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

17. F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, 1993.

18. E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings ot the 13th National Conference on Artificial Intelligence (AAAI)*, pages 1044–1049, 1996.

19. E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings ot the 16th National Conference on Artificial Intelligence (AAAI)*, pages 474–479, 1999.

APPENDIX: INSTRUCTIONS FOR THE RATING EXPERIMENT

You are helping us evaluate part of one component of a text-to-scene conversion system, specifically a part that attempts to make some "common sense" inferences about sentences.

You will be presented with a list of 90 sentences. Each of these sentences describes some action or situation. Associated with each sentence are three suggestions for answers to the following questions about the sentence:

- What room in a house does the described action or situation occur in?
- At what time of day does it occur?
- In what season does it occur?

The answers have been provided by an automatic procedure that attempts to predict the answers on the basis of the actions mentioned in the sentence. Your task is to decide if these predictions are correct. Specifically:

- If the prediction is correct — i.e. accords with your judgment — then leave the selection alone.
- If the prediction is clearly wrong then change it to what, in your view, is the correct answer.

In each case the correct answer is one of a selected group of answers (e.g. for season: **summer**, **winter**, **spring**, **autumn**), or **ANY**. **ANY** should be chosen if the action or situation does not seem to imply a particular location or time.

There may be some instances in which more than one, but not all of the answers are possible. For example something may be likely to take place at any time during the day, but not at night. In such cases you should be lenient with preselected answers that are in the acceptable set (i.e. don't change the selection), but if the preselected answer is not in the acceptable set, then choose **one member of the acceptable set** (rather than **ANY**) as your answer. For example if the provided answer is "night", and the activity could take place at any time during the day (but not at night), then select, say, "afternoon", or "morning".

For rooms, there is the additional option **NONE OF ABOVE**. This should be selected if in your view the action or situation must occur outside or in any event cannot be in one of the listed rooms.

You should try to base your judgments on what first comes to mind, rather than on deep introspection. For example, if the sentence is

John is washing his dog

and the first thing that comes to mind is that he must be in the bathroom, then that should be considered the correct answer. You may then reason that perhaps he has a tub of water in his living room, but you should avoid considering that as the correct answer.