Extracting aspects of determiner meaning from dialogue in a virtual world environment

Extracting aspects of determiner meaning from dialogue in a virtual world environment

Hilke Reckman, Jeff Orkin, and Deb Roy
MIT Media Lab
{reckman,jorkin,dkroy}@media.mit.edu

Abstract

We use data from a virtual world game for automated learning of words and grammatical constructions and their meanings. The language data are an integral part of the social interaction in the game and consist of chat dialogue, which is only constrained by the cultural context, as set by the nature of the provided virtual environment. Building on previous work, where we extracted a vocabulary for concrete objects in the game by making use of the non-linguistic context, we now target NP/DP grammar, in particular determiners. We assume that we have captured the meanings of a set of determiners if we can predict which determiner will be used in a particular context. To this end we train a classifier that predicts the choice of a determiner on the basis of features from the linguistic and non-linguistic context.

1 Introduction

Determiners are among those words whose meanings are hardest to define in a dictionary. In NLP, determiners are often considered 'stop words' that are not relevant for understanding the content of a document and should be removed before any interesting processing is done. On the other hand, it has been shown that children are sensitive to determiner choice already at a very early age, using these function words in figuring out what content nouns are intended to refer to. Meanings of determiners have been argued to include important pragmatic and discourse-related functions.

We have a corpus of dialogue that is grounded in a virtual environment. This means that in our data there is a relation between what people are saying and what they are doing, providing cues as to what they mean by the words and constructions they use. We have chosen to use a virtual world environment to collect data in, rather than a real world environment, because relatively rich virtual worlds are by now available that are able to provide an interesting level of grounding, whereas making sense of real world scenes using computer vision is still very challenging. In addition, this choice allows us to conveniently collect data online[1].

Although there exists a rich body of computational linguistics research on learning from corpus data, these corpora usually consist of text only. Only recently corpora that include non-linguistic context have started to be collected and used for grounded learning of semantics (Chen et al., 2010; Frank et al., 2009; Fleischman and Roy, 2005; Gorniak and Roy, 2005). This kind of work offers new and insightful perspectives on learning meanings of natural language words and constructions, based on the idea that our own knowledge of natural language meanings is grounded in action and perception (Roy, 2005), and that language is a complex adaptive system which evolves in a community through grounded interaction (e.g. Steels, 2003). So far the language in virtually grounded datasets has often been restricted to either descriptions or directives, so utterances can be paired fairly directly with the actions they describe. The interaction in our data is much freer. That means that it is more representative for the data that human learners get, and that our methods can be applied to a wider variety of data, possibly also to datasets

[1] Von Ahn and Dabbish (2004) were among the first to realize the potential of collecting human knowledge data online, in a game setup, collecting a large image-labeling corpus.

that have not been collected specifically for this purpose. A related project is KomParse (Klüwer et al., 2010). Piantadosi et al. (2008) developed a Bayesian model that learns compositional semantic meanings of different kinds of words, including quantifiers, but from completely artificial data.

Our research focuses on learning from data, rather than through interaction, though the latter may be possible in a later stage of the project. An example of a virtual world project where language is learned through interaction is 'Wubble World' (Hewlett et al., 2007). In the Give Challenge (Byron et al., 2009) a virtual world setup is used to evaluate natural language generation systems.

In previous work we have extracted words and multi-word expressions that refer to a range of objects that are prominent in our virtual environment (Reckman et al., 2010). Now we investigate if aspects of determiner meaning can be learned from this dataset. The extracted knowledge of nouns makes the learning of determiners possible. We study what factors contribute to the choice of the determiner and how they relate to each other, by training a decision tree classifier using these factors as features. The decision tree provides insight in which features are actually used, in which order, and to which effect. The accuracy of the resulting classifier on a test set should give us an impression of how well we understand the use of the different determiners. Although one may argue that this study is about use rather than about meaning, we take it that meaning can only be learned through use, and it is meaning that we are ultimately interested in. One of the overarching questions we are concerned with is what knowledge about language and how it works is needed to extract knowledge about constructions and their meanings from grounded data. Practically, a computational understanding of determiners will contribute to determining the reference of referential expressions, particularly in situated dialogue, and to generating felicitous referential expressions (cf. Belz et al., 2010).

We first introduce our dataset. Then we discuss the automated extraction of determiners. Subsequently, we motivate the features we use, present our classifier experiments, and discuss the results.

2 Data: The Restaurant Game

Orkin and Roy (2007) showed in The Restaurant Game project that current computer game technology allows for simulating a restaurant at a high level-of-detail, and exploiting the game-play experiences of thousands of players to capture a wider coverage of knowledge than what could be handcrafted by a team of researchers. The restaurant theme was inspired by the idea of Schank and Abelson (1977), who argued that the understanding of language requires the representation of common ground for everyday scenarios. The goal is automating characters with learned behavior and dialogue. The ongoing Restaurant Game project has provided a rich dataset for linguistic and AI research. In an online two-player game humans are anonymously paired to play the roles of customers and waitresses in a virtual restaurant (http://theRestaurantGame.net). Players can chat with open-ended typed text, move around the 3D environment, and manipulate 47 types of interactive objects through a point-and-click interface (see figure 1). Every object provides the same interaction options: pick up, put down, give, inspect, sit on, eat, and touch, but objects respond to these actions in different ways. The chef and bartender are hard-coded to produce food items based on keywords in chat text. A game takes about 10-15 minutes to play. Everything players say and do is logged in time-coded text files on our servers. Although player interactions vary greatly, we have demonstrated that enough people do engage in common behavior that it is possible for an automatic system to learn statistical models of typical behavior and language that correlate highly with human judgment of typicality (Orkin and Roy, 2007).

Over 10.000 games have been collected. The dialogue is grounded in two (partially overlapping) ways. Not only is there a simulated physical environment with objects that can be manipulated in various ways, but also social patterns of recurring events provide an anchor for making sense of the dialogue. Previous research results include a first implementation of a planner that drives AI characters playing the game (Orkin and Roy, 2009).

The intuition is that a human student of English starting from scratch (but with some common sense knowledge about restaurants), could learn quite a bit of English from studying the Restaurant Game episodes; possibly enough to play the game. We try to computationally simulate such a learning process.

Figure 1: Screen-shots from The Restaurant Game, from left to right: third-person perspective, waitress's perspective with dialogue, menu for interacting with objects.

3 Extracting nouns

Previously, we extracted a vocabulary of referring expressions for a set of concrete objects, based on which words and phrases have the highest relative frequency in the contexts in which the objects are used (see figure 2). We extracted words and phrases that can refer to the food and drink items on the restaurant's menu, the menu, and the bill, and some other items. These expressions represent the core nominal phrases in the game. We will use these expressions as a starting point to extract determiners and nominal modifiers. We restrict ourselves to the ordered food and drink items, the menu and the bill, expecting that these show a somewhat uniform and interesting behavior, as they are the objects that can appear and disappear during the course of a game.

food type	referring expressions
SOUP	'soup' 'vegetable soup' 'soup du jour' 'soup de jour'
SALAD	'salad' 'cobb salad'
SPAGHETTI	'spaghetti' 'spaghetti marinara'
FILET	'steak' 'filet' 'filet mignon'
SALMON	'salmon' 'grilled salmon'
LOBSTER	'lobster' 'lobster thermador'
CHEESECAKE	'cheesecake' 'cheese' 'cake' 'cherry cheesecake' 'cheese cake'
PIE	'pie' 'berry pie'
TART	'tart' 'nectarine tart'

drink type	referring expressions
WATER	'water'
TEA	'tea'
COFFEE	'coffee'
BEER	'beer'
REDWINE	'red' 'wine' 'red wine'
WHITEWINE	'white' 'white wine'

item type	referring expressions
MENU	'menu'
BILL	'bill' 'check'

Figure 2: Extracted referring expressions for relevant items.

The referring expressions for these object types have been extracted in an unsupervised manner making use of the relative frequency of words and phrases in the context of the objects being used. Words, bigrams and trigrams were validated against each other with the use of one threshold. For more detail see (Reckman et al., 2010).

4 Extracting determiners

Extracting determiners totally unsupervised is a non-trivial task. Attempts to use the existing fully unsupervised grammar induction algorithm ADIOS (Solan et al., 2005) did not give us the results we were hoping for. Instead, we decided to make use of the knowledge of nouns that we already have and target

determiners directly, rather than having to induce a full grammar. In future work we will look into using alternative grammar induction systems, for a wider range of learning tasks.

We first narrowed down our search space by collecting words that are positively associated with the position directly to the left of the nominal expression above a high recall, low precision threshold (phi=0.01)[2]. This should favor determiners and other nominal modifiers over, for example, verbs.

We expect determiners to appear with a wider range of different nouns than adjectival modifiers do. Especially in this restricted domain, adjectives are more likely to be restricted to specific object types. We consider pre-nominal terms that are general enough to appear with more than 5 different objects (out of 17) to be determiner candidates. We also check that our candidates can be preceded by an utterance boundary.

The word *the* is most strongly associated with the relevant position, combines with most different nouns, and can occur as only element between a boundary and a noun. We therefore assume that at least *the* is a determiner. We order the other candidates according to their similarity to *the*, measured as the cosine distance in a vector-space, with their two words to the left and to the right as dimensions. We accept words as determiners in order of similarity to *the*, starting with the most similar word, after checking that they are in complementary distribution with all of the already accepted words, i.e. that the word does not occur adjacent to any of those. This gives us the following determiners: *the*, *my*, *your*, *some*, *a*, *another*, *our*, *one*, *ur*, *two*, *2*.[3]

We can then identify adjectival modifiers by looking at what occurs between determiners and nouns. By checking what else these modifiers can be preceded by (that is also in complementary distribution with known determiners), we can do another round of determiner search, and that lets us add *any* to our list. As nouns can also be immediately preceded by an utterance boundary, we establish that the determiner position is not obligatorily filled.

Of course this is not a complete set of determiners, but they appear to be the most prominent ones in the game. Real quantifiers are relatively rare and that is to be expected, given the setting. Perhaps more surprisingly, *this* and *that* are not associated with the left-of-noun position. It turns out that they are not used very frequently as determiners in the game, and much more as pronouns. In future work we will extract pronouns, by looking for single words that have a distribution that is similar to the distribution of full noun phrases with a determiner.

In the process of extracting determiners, we also extract adjectives and modifiers such as *glass of*. With little extra effort we can build a vocabulary of these as well, including information as to which nouns they are associated with. Their meanings, however, are in most cases not sufficiently grounded in the game to be understood. We may in a more advanced stage of the project be able to figure out that the adjective *free* makes the item less likely to appear on the bill, but the meaning of *hot* will always remain unclear, as temperature is not modeled in the game. Finding words associated with the position to the left of specific nouns can also help us further improve our vocabulary of referring expressions, for example by identifying *veg* and *veggie* as alternatives for *vegetable* in *vegetable soup*[4].

We took a shortcut by directly targeting the position left of the noun. This involves language-specific knowledge about English. To make this method applicable to different languages and only use very general knowledge at the start, we would first have to find out what the position of the determiner is. This may be to the right of the noun or affixed to it. Not all languages have articles, but we can expect determiners like *my*, *your*, *another* etc. to occur either adjacent to[5], or morphologically expressed on the noun[6]. In previous work we have shown how a construction for coordination can be extracted (Reckman

[2]The phi-score is a chi-square based association metric. Manning and Schütze (2000) argue that such metrics are suitable to quantify collocational effects. We also used it in extracting the referring expressions.

[3]For the experiments we replace *ur* by *your*, and *2* by *two*. We assume this could in principle be done automatically, although especially in the latter case this is not trivial.

[4]We do already have a list of spelling variants for all the terms, but *veg* and *veggie* were too different from the canonical form to get through the edit-distance filter

[5]Obviously we do not catch floating quantifiers this way. We might catch their non-floating counterparts and then discover that they occur in other positions as well.

[6]Several unsupervised morphological analyzers have been developed, which should in principle be run in an early stage of learning. For English however, the only interesting morphology at play here is plural formation.

et al., 2010). Coordination, to our knowledge, occurs in all languages and this is probably a feature of general human cognition, so it makes sense to assume it exists in a language and look for it in the data. It can then be used as a probe on structure. Categories that are grammatically intimately connected to nouns are more likely to be repeated in a coordination involving two nouns. If we look at our English data, for example, we see that a lot more material tends to occur between *and* and the second noun-conjunct, than between the first noun-conjunct and *and*, which suggests that things that are grammatically close to the noun occur to the left of it.

5 Features

In this section we motivate the features we will use. To capture the full meaning of determiners, we would probably have to model the mental states of the players. However, what we aim at here is a preliminary understanding of determiners as a step towards the understanding of full sentences, and the resolution of NP reference and co-reference, which would be prerequisites for any serious modeling of mental states. So we are interested in what can be learned from directly observable features. The features are theoretically motivated, and reflect the nature of the referent, whether the referent has been mentioned before, whether the referent is present, and who the speaker and addressee are.

The first feature is object type. There are 17 different objects that we take into account: BEER, BILL, CHEESECAKE, COFFEE, FILET, LOBSTER, MENU, PIE, REDWINE, SALAD, SALMON, SOUP, SPAGHETTI, TART, TEA, WATER, and WHITEWINE. We expect this feature to matter, because in a restaurant situation one usually orders '*the* spaghetti', but '*a* beer'. This may be to some extent dependent on what is on the menu, but not completely. Regardless of what is on the menu, ordering '*the* Heineken' seems to be more unusual than ordering '*the* Merlot'. This may mean that our data is not entirely representative of the general case, because of our restaurant setting. However, it cannot be excluded that similar effects play a role in other settings, too. There is of course the effect of mass versus count nouns, too, but this may be a bit masked, because of unit expressions like *glass of*. We chose to not include these unit expressions as a feature, because the decision to use such modifiers can be considered part of the decision on which determiner to use. So using the modifier as a feature, would be giving away part of the solution to the determiner-choice problem.

The second feature captures the notion of discourse-old versus discourse-new. We distinguish between cases where an object of a particular type is mentioned for the first time, and where it has already been mentioned before. In the latter case, we take it that the discourse referent has already been introduced. The expected effect is that first mentions tend to be indefinite.[7] This is only an approximation, because sometimes a second object of the same type is introduced and we do not resolve the reference of our instances.

The third and fourth features incorporate present versus future presence of the object, plus the position of the utterance with respect to the central action involving the object. We keep track of the previous and following action in which the object is involved. Actions of interest are restricted to the appearance of the object and and its central action: 'eating' for food and drink items, 'looking at' for the menu, and 'paying' for the bill. Being involved in such an action also implies presence. Other intervening actions are ignored. The features are 'preceding_action' and 'following_action', and the values are 'appearance', 'main_action', and 'none'. We expect indefinites before appearance, when the object is not yet present. Note that these features rely entirely on non-linguistic context.

The fifth and sixth features identify speaker and addressee. The speaker can be the customer or the waitress. For the addressee the relevant distinction is whether the staff (chef and bartender) are addressed or not. We expect a tendency of the waitress using *your* when talking to the customer, and of the customer using *my* more often. We expect more indefinites or absence of a determiner when the staff is spoken to. These features are central to dialogue, and may reveal differences between the roles.

[7] This is a typical feature for languages that have articles, and may be expressed through other means in other languages.

6 Experiments

We use the decision tree classifier from the Natural Language ToolKit for Python (Loper and Bird, 2002) and train and test it through 10-fold cross-validation on 74304 noun phrases from 5000 games, 23776 of which actually have determiners. The noun phrases used all contain nouns that can refer to the selected objects, though we cannot guarantee that they were intended to do so in all cases. In fact, we have seen examples where this is clearly not the case, and for example *filet*, which normally refers to the FILET object, is used in the context of salmon. This means that there is a level of noise in our data.

The instances where the determiner is absent are very dominant, and this part of the data is necessarily noisy, because of rare determiners that we've missed[8], and possibly rather heterogeneous, as there are many reasons why people may choose to not type a determiner in chat. Therefore we focus on the experiments where we have excluded these cases, as the results are more interesting. We will refer to the data that excludes instances with no determiner as the **restricted dataset**. When instances with no determiner are included, we will talk about the **full dataset**.

6.1 Baselines

In the experiments we compare the results of using the features to two different baselines. The simplest baseline is to always choose the most frequent determiner. For the instances that have overt determiners, the most frequent one is *the*. Always choosing *the* gives us a mean accuracy of 0.364. If we include the instances with no overt determiners, that gives us a much higher baseline of 0.680, when the no determiner option is always chosen. We call this the **simple baseline**.

The second baseline is the result of using only the object feature, and forms the basis of our experiments. We call this the **object-only baseline**. On the restricted dataset the resulting classifier assigns the determiner *a* to the objects BEER, COFFEE, PIE, REDWINE, SALAD, TEA, WATER, and WHITEWINE, and the determiner *the* to BILL, CHEESECAKE, FILET, LOBSTER, MENU, SALMON, SOUP, SPAGHETTI, and TART. This yields a mean accuracy of 0.520, which is a considerable improvement over the simple baseline that is relevant for this part of the data. If we look at the confusion matrix in figure 3 that summarizes the results of all 10 object-only runs we see that the objects' preferences for definite versus indefinite determiners are also visible in the way instances with determiners other than *the* and *a* are misclassified. Instances with definite determiners are more often classified as *the*, and indefinites as *a*.

	a	another	any	my	one	our	some	the	two	your
a	<4984>	.	.	.	.	.	.	2912	.	.
another	608	<.>	.	.	.	.	.	76	.	.
any	56	.	<.>	.	.	.	.	24	.	.
my	238	.	.	<.>	.	.	.	742	.	.
one	354	.	.	.	<.>	.	.	241	.	.
our	28	.	.	.	.	<.>	.	178	.	.
some	1109	.	.	.	.	.	<.>	438	.	.
the	1270	.	.	.	.	.	.	<7383>	.	.
two	191	.	.	.	.	.	.	58	<.>	.
your	805	.	.	.	.	.	.	2075	.	<.>

Figure 3: Confusion matrix for the object-only baseline.

On the full dataset, the classifier assigns *the* to instances of BILL and MENU and no determiner to everything else, reflecting the count/mass distinction, and resulting in a mean accuracy of 0.707. This is also a statistically significant improvement over its baseline, but much less spectacular. The definite/indefinite distinction that we saw with the restricted dataset, does not really emerge here.

[8]It is also hard to reliably recognize misspelled determiners as determiners tend to be very short words.

6.2 Adding the other features

In the core experiments of this paper we always use the object feature as a basis and measure the effect of adding the other features, separately and in combination. All differences reported are significant, unless stated otherwise. The table in figure 5 at the end of the section summarizes the results.

If we add the feature of whether the item has been mentioned before or not, we get more indefinites, as was to be expected. On the restricted dataset, the MENU, PIE, and TART objects get *a* if not mentioned previously, and *the* otherwise. The mean accuracy is 0.527, which is a statistically significant improvement over the object-only baseline (the improvement is consistent over all 10 runs), but it seems rather small, nevertheless. (Using the discourse feature without the object feature gives a score of 0.377.) Adding information as to whether the customer has seen the menu does not make any difference. On the full dataset the discourse feature matters only for MENU, which gets *a* if not previously mentioned. The mean accuracy is 0.709.

If, instead, we add the action features we get a somewhat more substantial improvement for the restricted dataset; a mean accuracy of 0.561. We also get a wider range of determiners: *your* tends to be chosen after appearing and before eating, *another* after eating, and *a* between no action and appearing. The order in which the following and preceding action features are applied by the classifier differs per object. (The action features without the object feature give a mean accuracy score of 0.427.) For the full dataset the mean accuracy is 0.714, again a consistent, but marginal improvement. However, *a*, *the* and *your* are the only determiners used, in addition to the no determiner option.

Adding the speaker and addressee features to the object feature base gives the classifier a better grip on *your*. More indefinites are used when the staff is addressed, *your* when the customer is spoken to. However, *my* is still not picked up. The speaker and addressee features are used in both orders. The mean accuracy is 0.540, which is better than with the discourse feature, but worse than with the action features. (The speaker and addressee features without the object feature give a mean accuracy score of 0.424.) In the case of the full dataset, the new features are barely used, and there is no consistent improvement over the different runs. The mean accuracy is 0.711.

If we combine the action features and speaker/addressee features on top of the object feature basis, we see a substantial improvement again for the restricted dataset. The mean accuracy is 0.592. Finally, we get some cases of *my* being correctly classified, and also *your* is correctly classified significantly more often than in the previous experiments. The object feature always comes first in the decision tree. For the other features, all relative orders are attested. Adding the 'previously-mentioned' feature to this combination (see also figure 4) improves this result a little bit more, to a mean accuracy of 0.594, although we can expect the information contained in it to have a large overlap with the information in other features, for example, items mentioned for the first time will typically not have appeared yet.

	a	another	any	my	one	our	some	the	two	your
a	<5732>	163	1	11	20	.	70	1773	1	125
another	175	<350>	.	2	.	.	48	70	.	39
any	44	4	<.>	.	.	.	2	29	.	1
my	154	19	.	<9>	.	.	20	765	.	13
one	437	20	.	2	<16>	.	4	70	.	46
our	29	1	.	.	.	<.>	1	161	.	14
some	881	48	.	6	3	.	<114>	421	.	74
the	1332	74	.	33	8	.	34	<6131>	.	1040
two	191	10	.	2	.	.	1	45	<.>	.
your	218	88	.	.	.	.	20	781	.	<1773>

(row = reference; col = test)

Figure 4: Confusion matrix for the object, action, speaker/addressee and discourse features combined.

6.3 Linguistic context and dialogue acts

It will be part of future research to distinguish the different dialogue acts that the nominal phrases that we studied can be part of. Identifying the 'task' that an expression is part of may have a similar effect. Tasks of the type 'customer gets seated', 'waitress serves food', 'customer eats meal', etc. are annotated for supervised learning, and may consist of several actions and utterances (Orkin et al., 2010).

To give an indication that the dialogue act that an expression is part of may be informative as to the correct choice of the determiner, we have done an extra experiment, where we have used the word before and the word after the DP as features. This gives a tremendous amount of feature values, which are not very insightful, due to the lack of generalization, and are a near guarantee for over-fitting. However, it does yield an improvement over using the object-only baseline. Moreover, the preceding word and following word features are now applied before the object feature. The mean accuracy in this experiment was 0.562, which is comparable to the experiment with object and action features. At the same time we get a wider range of determiners than we have had before, including some correctly classified instances of *our*. On the full dataset we even get a higher accuracy score than in any of the other experiments: 0.769, also with a much wider range of determiners. We suspect that this local linguistic context gives quite good cues as to whether the expression is part of a proper sentence or not, and that in the former case an overt determiner is much more likely[9]. The results of all experiments are summarized in figure 5.

	restricted	full
simple baseline	0.364	0.680
object-only baseline	0.520	0.707
object + discourse	0.527	0.709
object + action	0.561	0.714
object + speaker	0.540	0.711
object + action + speaker	0.592	0.721
object + action + speaker + discourse	**0.594**	0.721
object + surrounding words	0.562	**0.769**

Figure 5: Summary of the testing results.

7 Discussion

Maybe the most surprising outcome is that the object type turns out to be the main factor in choosing the determiner in this virtual restaurant setting. It would beinteresting to see this reproduced on the data of two new games that are currently being developed, with novel scenarios, locations and objects. At the same time, it is a strength of our approach, that we can simulate a specific setting and capture its ideosyncrasies, learning domain-specific aspects of language, and hopefully eventually learn what generalizes across different scenarios.

For the restricted dataset we see that, consistently, indefinites are mostly misclassified as *a*, and definites mostly as *the*. If we evaluate only for definiteness, we get a mean accuracy of 0.800 for the case with all features combined. We could distinguish these two classes of determiners on the basis of the similarity of each determiner to the two dominant types. It is, however, the object feature that seems to be mainly responsible for the gain in definiteness accuracy with respect to the simple baseline.

It is unsurprising that we haven't learned much about *one* and *two*, except that they pattern with indefinites, as we haven't included features that have to do with the number of objects. There actually are more numerals that appear in the game, but did not make it into our list of determiners, because they did not occur with enough different objects. In the general case, we are doubtful that numerals are sufficiently grounded in this game for their exact meanings to be learned. It may however be possible to learn a one-two-many kind of distinction. This would also involve looking into plural morphology, and remains for future research.

[9] We have observed that in several games people tend to just sum up food items, without embedding them in a sentence.

We also haven't learned anything about *our*, except that it patterns with definites. It is not quite clear what kind of features would be relevant to *our* in this setting.

For the possessive pronouns *your* and *my* we have learned that one tends to be linked to the waitress as a speaker (and the customer as addressee) and the other to the customer. It will be challenging to reach an understanding that goes deeper than this[10]. The range of interactions in the game may be too limited to learn the meanings of all determiners in their full generality.

While we have treated *a* and *another* as different determiners, we have included cases of *some more* under *some*. It may be worthwhile to include *some more* (and perhaps *any more* and *one more* as well) as a separate determiner. However, our best classifier so far still cannot distinguish between *a* and *another* very well.

The experiments with linguistic context suggest that dialogue act may make for an additional, powerful, albeit indirect, feature. The fact that it helps to know when the main action involving the object took place, rather than just its appearance, may also be taken to point in the same direction, as people tend to say different kinds of things about an object before and after the main action.

Using a classifier seems to be a reasonable way of testing how well we understand determiners, as long as our features provide insight. Although there is still a lot of room for improvement, there is likely to be a ceiling effect at some point, because sometimes more than one option is felicitous. We also have to keep in mind that chat is likely to be more variable than normal written or spoken language.

8 Conclusion

We have carried out an exploratory series of experiments, to see if meanings of determiners, a very abstract linguistic category, could be learned from virtually grounded dialogue data. We have trained a classifier on a set of theoretically motivated features, and used the testing phase to evaluate how well these features predict the choice of the determiner.

Altogether, the results are encouraging. If we exclude instances with no determiner we reach an accuracy of 0.594 over a baseline of 0.364. The features that identify the dialogue participants and surrounding actions, including appearance, play an important role in this result, even though the object type remains the main factor. A clear dichotomy between definite and indefinite determiners emerges. The results for the complete dataset are a bit messier, and need more work.

In future work we will identify utterance types, or dialogue acts, that also rely on surrounding actions and on the speaker and addressee. We will also look into resolving reference and co-reference.

Acknowledgments

This research was funded by a Rubicon grant from the Netherlands Organisation for Scientific Research (NWO), project nr. 446-09-011.

References

Belz, A., E. Kow, J. Viethen, and A. Gatt (2010). Generating referring expressions in context: The GREC task evaluation challenges. In *Empirical Methods in Natural Language Generation*, pp. 294–327. Springer.

Byron, D., A. Koller, K. Striegnitz, J. Cassell, R. Dale, J. Moore, and J. Oberlander (2009). Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation*, pp. 165–173. ACL.

Chen, D., J. Kim, and R. Mooney (2010). Training a Multilingual Sportscaster: Using Perceptual Context to Learn Language. *Journal of Artificial Intelligence Research 37*, 397–435.

[10]For their personal pronoun counterparts *you* and *I* we might stand a better chance.

Fleischman, M. and D. Roy (2005). Why verbs are harder to learn than nouns: Initial insights from a computational model of intention recognition in situated word learning. In *27th Annual Meeting of the Cognitive Science Society, Stresa, Italy*.

Frank, M., N. Goodman, and J. Tenenbaum (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science 20*(5), 578.

Gorniak, P. and D. Roy (2005). Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the 7th international conference on Multimodal interfaces*, pp. 143. ACM.

Hewlett, D., S. Hoversten, W. Kerr, P. Cohen, and Y. Chang (2007). Wubble world. In *Proceedings of the 3rd Conference on Artificial Intelligence and Interactive Entertainment*.

Klüwer, T., P. Adolphs, F. Xu, H. Uszkoreit, and X. Cheng (2010). Talking NPCs in a virtual game world. In *Proceedings of the ACL 2010 System Demonstrations*, pp. 36–41. ACL.

Loper, E. and S. Bird (2002). NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pp. 70. ACL.

Manning, C. and H. Schütze (2000). *Foundations of statistical natural language processing*. MIT Press.

Orkin, J. and D. Roy (2007). The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development 3*(1), 39–60.

Orkin, J. and D. Roy (2009). Automatic learning and generation of social behavior from collective human gameplay. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 385–392. International Foundation for Autonomous Agents and Multiagent Systems.

Orkin, J., T. Smith, H. Reckman, and D. Roy (2010). Semi-Automatic Task Recognition for Interactive Narratives with EAT & RUN. In *Proceedings of the 3rd Intelligent Narrative Technologies Workshop at the 5th International Conference on Foundations of Digital Games (FDG)*.

Piantadosi, S., N. Goodman, B. Ellis, and J. Tenenbaum (2008). A Bayesian model of the acquisition of compositional semantics. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*. Citeseer.

Reckman, H., J. Orkin, and D. Roy (2010). Learning meanings of words and constructions, grounded in a virtual game. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS)*.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence 167*(1-2), 170–205.

Schank, R. and R. Abelson (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates Hillsdale, NJ.

Solan, Z., D. Horn, E. Ruppin, and S. Edelman (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America 102*(33), 11629.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in cognitive sciences 7*(7), 308–312.

Von Ahn, L. and L. Dabbish (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326. ACM.