Semantic Annotation for Generation: Issues in annotating a corpus to develop and evaluate discourse entity realization algorithms

Massimo Poesio
University of Edinburgh
HCRC and Informatics
{Massimo.Poesio}@ed.ac.uk

Abstract

We are annotating a corpus with information relevant to discourse entity realization, and especially the information needed to decide which type of NP to use. The corpus is being used to study correlations between NP type and certain semantic or discourse features, to evaluate hand-coded algorithms, and to train statistical models. We report on the development of our annotation scheme, the problems we have encountered, and the results obtained so far.

1 MOTIVATIONS

The goal of the GNOME project is to develop NP generation algorithms that can be used by real systems, with different architectures, and operating in realistic domains. As part of the project, we have been annotating a corpus with the syntactic, semantic and discourse information that is needed for different subtasks of NP realization, including the task of deciding on the most appropriate NP type to be used to realize a certain discourse entity (proper name, definite description, pronoun, etc.), and the task of organizing the additional information to be expressed with that discourse entity. We are using the annotated corpus to extract information useful to the development of hand-coded algorithms for the subtasks of NP realization we are focusing on, to develop statistical models of these subtasks, and to evaluate both types of algorithms. Conversely, we have been using the results of this evaluation to verify the completeness of our annotation scheme and to identify modifications. The annotation scheme used in our first corpus annotation exercise was discussed in (Poesio et al., 1999b); in this paper we present the modified annotation scheme that we developed as a result of that preliminary work, and discuss the problems we encountered when trying to annotate semantic and discourse information.

2 APPLICATIONS AND DATA

The systems we are working with are the ILEX system developed at HCRC, University of Edinburgh (Oberlander et al., 1998),[1] and the ICONOCLAST system (Scott et al., 1998), developed at ITRI, University of Brighton. The ILEX system generates Web pages describing museum objects on the basis of the perceived status of its user's knowledge and of the objects she previously looked at; ICONOCLAST supports the creation of pharmaceutical leaflets by means of the WYSIWYM technique in which text generation and user input are interleaved.

The corpus we have collected for GNOME includes texts from both the domains we are studying. It contains texts in the museum domain, extending the corpus collected by the SOLE project (Hitzeman et al., 1998); and texts from the corpus of patient information leaflets collected for the ICONOCLAST project. The initial GNOME corpus (Poesio et al., 1999b) consisted of two subsets of about 1,500 NPs each; since then, the corpus has been extended and currently includes about 3,000 NPs in each domain. We are also adding texts from a third domain, tutorial dialogues.

3 DEVELOPING A SCHEME FOR NP REALIZATION

The traditional approach to surface realization in NLG (as exemplified, say, by NIGEL / KPML (Henschel et al., 1999)) assumes (systemic functional) grammars that make decisions on the basis of the answer to queries asked to the knowledge base and discourse model. Typical examples of such queries are:

- whether a given discourse entity is IDENTIFIABLE;

[1] The latest version of the system can be found at `http://www.cstr.ed.ac.uk/cgi-bin/ilex.cgi`.

- whether the object denoted is GENERIC or not;
- whether that entity is IN FOCUS, or more generally what is its ACCESSIBILITY (Gundel et al., 1993)
- what is the ONTOLOGICAL STATUS of the object, i.e., its position in a taxonomy.

These systems have typically been used only by their developers, or by researchers working in close collaboration with them. In order to make them more generally usable, three questions have to be addressed. The first question is whether anybody other than the developers of these grammars can understand queries such as those just listed enough to implement them in their systems. The second is whether real systems have enough information to answer these queries, or whether instead approximations have to be implemented. The final question is how well the implementation is going to perform, especially if only approximations are implemented.

In GNOME we have been studying these questions by means of corpus annotation studies. We have been trying to identify which of the queries used by systems such as KPML for NP realization can be generally understood by asking subjects to annotate the NPs in our corpus with the information needed to answer these queries, and we have then used the resulting annotation to train statistical models to evaluate the completeness of a given set of features. We use to measure agreement the K statistic discussed by Carletta (1996). A value of K between .8 and 1 indicates good agreement; a value between .6 and .8 indicates some agreement.

4 SEMANTIC AND DISCOURSE FEATURES THAT MAY AFFECT NP TYPE DETERMINATION

Even if in this first phase we focused on realizing discourse entities only, we still need to know for each NP in the corpus its semantic type. Noun phrases appear in a text as the realization of at least three different types of logical form constituents:

- **terms**, which include referring expressions, as in *Jessie M. King* or *the hour pieces here* , but also non-referring terms such as *jewelry* or *different types of creative work*. Terms are called DISCOURSE ENTITIES in Discourse Representation Theory.
- **quantifiers**, as in *quite a lot of different types of creative work* or *nearly every day*
- **nominal predicates**, such as *an illustrator* in *She was an illustrator*.

Noun phrases can be **coordinated**, as in *The patches also contain oestradiol and norethisterone acetate* or *the inventory gives neither the name of the maker nor its original location*; we finesse the many issues raised by coordination by assuming a fourth type of logical form objects, **coordinations**.

Two features generally acknowledged to play an important role in determining the type of the NP to be used to realize a discourse entity are COUNTABILITY and GENERICITY. These features are especially important when bare-NPs are going to be used. One of the conditions under which (singular) bare NPs are used is when the object denoted is mass (cfr. **a gold/a jewel* vs. *gold/*jewel*); the other is when the NP is used to express a generic reference, as in *The cabinets de curiosites contained natural specimens such as shells and fossils*.

Much work on NP generation has been devoted to studying the discourse factors that determine whether a given discourse entity should be realized by a definite or an indefinite NP (Prince, 1992; Loebner, 1987; Gundel et al., 1993). Among the discourse properties of a discourse entity claimed to affect its form are

- Whether it is discourse new or old (Prince, 1992): e.g., a new jewel would be introduced by means of the indefinite *a jewel*, whereas for an already mentioned one the definite description *the jewel* would be used. This simple notion of familiarity was refined by Prince herself as well by Gundel *et al*. (Gundel et al., 1993).
- Whether it's hearer-new or hearer-old (Prince, 1992).
- Whether it is referring to an object in the visual situation or not: if so, a demonstrative NP may be used, as in *this jewel*.
- Whether it's currently highly salient or not, which may prompt the use of a pronoun. Properties that have been claimed to affect the salience of a discourse entity include: whether it's the current CENTER (CB) or not (Grosz et al., 1995), or more generally whether that entity is the TOPIC of the current discourse (Reinhart, 1981; Garrod and Sanford, 1983); its grammatical function; whether it's animated or

not; its role; its proximity. (For a discussion of the effect of these and other factors on salience see (Poesio and Stevenson, To appear)).

According to Loebner (Loebner, 1987), the distinguishing property of definites is not familiarity (a discourse notion), but whether or not the predicate denoted by the head noun is functional or, more generally, UNIQUE. This seems to be the closest formal specification of the notion of 'identifiability' used in KPML.

5 THE ANNOTATION SCHEME

Our first scheme, and the results we obtained with it, are discussed in (Poesio et al., 1999b). We are currently in the process of reannotating the corpus from scratch according to a new annotation scheme developed to address the limitations of the scheme discussed there (reliability and/or incompleteness of information). The new scheme also includes information to study another aspect of NP realization, NP modification; this aspect of the new annotation won't be discussed here. For reasons of space, only a brief discussion is possible - in particular, we won't be able to discuss in detail the instructions given to annotators; the complete instructions are available at `http://www.hcrc.ed.ac.uk/~gnome/anno_manual.html`.

Markup Language

Our annotation scheme is XML-based. The basis for our annotation are a rather minimal set of layout tags, identifying the main divisions of texts, their titles, figures, paragraphs, and lists. Also, as a result of the reliability studies discussed below and of our first annotation effort, we decided to also mark up units of text that may correspond to rhetorical units in our second annotation, using the tag ⟨unit⟩.

An important feature of the scheme is that the information about NPs is split among two XML elements, as in the MATE scheme for coreference (Poesio et al., 1999a). Each NP in the text is tagged with an ⟨ne⟩ tag, as follows:

(1)
```
<ne ID="ne07" ...  >
Scottish-born, Canadian based jew-
eller,
Alison Bailey-Smith</ne>
...
<ne ID="ne08"> <ne ID="ne09">Her</ne>
materials</ne>
```

the instructions for identifying the ⟨ne⟩ markables are derived from those proposed in the MATE project scheme for annotating anaphoric relations (Poesio et al., 1999a), which in turn were derived from those proposed by Passonneau (Passonneau, 1997) and in MUC-7 (Chinchor and Sundheim, 1995).

Anaphoric relations are annotated by means of a separate ⟨ante⟩ element specifying relations between ⟨ne⟩s, also as proposed in MATE. An ⟨ante⟩ element includes one or more ⟨anchor⟩ element, one for each plausible antecedent of the current discourse entity (in this way, ambiguous cases can be marked). E.g., the anaphoric relation in (1) between the possessive pronoun with `ID ="ne09"` and the proper name with `ID ="ne07"` is marked as follows:

(2)
```
<ante current="ne09">
 <anchor ID="ne07" rel="ident" ... >
</ante>
```

(Discourse) Units

One difference between the annotation scheme we are using and the one discussed in (Poesio et al., 1999b) is that the problems we encountered trying to annotate centering information, proximity, and grammatical function (see also below) led us to mark up sentences and potential rhetorical units / centering theory utterances before marking up certain types of information about NPs such as grammatical function. The instructions for marking up units were in part derived from (Marcu, 1999); for each ⟨unit⟩, the following attributes were marked:

- utype: whether the unit is a main clause, a relative clause, appositive, a parenthetical, etc.
- verbed: whether the unit contains a verb or not.
- finite: for verbed units, whether the verb is finite or not.
- subject: for verbed units, whether they have a full subject, an empty subject (expletive, as in *there* sentences), or no subject (e.g., for infinitival clauses).

The agreement on identifying the boundaries of units was K = .9; the agreement on features was follows:

Attribute	K Value
utype	.76
verbed	.9
finite	.81
subject	.86

This part of the annotation has now been completed. The main difficulties we observed had to do with assigning an utterance type to parenthetical sentences.

NEs

After marking up units as discussed above, all NPs are marked up, together with a number of attributes. During our first round of experimentation we found that marking 'topics' in general was too difficult (K=.37), as was marking up thematic roles (K=.42); so although we haven't completely abandoned the idea of trying to annotate this information, in this second round we concentrated on improving the reliability for the other attributes. A few other attributes used in the previous scheme were dropped because they could be inferred automatically: among these are the feature disc specifying whether the discourse entity is discourse-new or discourse-old (redundant once antecedent information was marked up) and the feature cb used to mark whether the discourse entity is the current CB (Grosz et al., 1995) (which could be automatically derived from the information about grammatical function and units). We separated off information about the logical form type of an NP (quantifier, term, etc) from the information about genericity. Finally, new attributes were introduced to specify information which we found missing on the basis of our first evaluation: in particular, we decided to annotate information about the abstractness or concreteness of an object, and about its semantic plurality or atomicity. The revised list of information annotated for each NP includes:

- The output feature, cat, indicating the type of NP (e.g., bare-np, the-np, a-np).
- The other 'basic' syntactic features, num, per, and gen (for GENder).
- A feature gf specifying its grammatical function;
- The following semantic attributes:
  - ani: whether the object denoted is animate or inanimate
  - count: whether the object denoted is mass or count
  - lftype: one of `quant,term,pred,coord`
  - generic: whether the object denoted is a generic or specific reference
  - onto: whether the object denoted is concrete, an event, a temporal reference, or another abstract object
  - structure: whether the object denoted is atomic or not
- The following discourse attributes:
  - deix: whether the object is a deictic reference or not
  - loeb: whether the description used allows the reader to characterize the object as functional in the sense of Loebner (i.e., whether it denotes a single object, as in *the moon*, or at least a functional concept, like *father*)

A number of NP properties (e.g., familiarity) can be derived from the annotation of anaphoric information (below); in addition, a few properties of NPs are automatically derived from other sources of information - e.g., the type of layout element in which the NP occurs (in titles, bare-nps are often used) and whether a particular NP has uniquely distinguishing syntactic features in a given unit. All of these features can be annotated reliably, except for genericity; the results that we do have are as follows:

Attribute	K Value
cat	.9
gen	.89
num	.84
per	.9
gf	.85
ani	.91
count	.86
lftype	.82
onto	.80
structure	.82
deix	.81
loeb	.80

(One interesting point to note here is that agreement on lftype is actually quite high (90%), but because `TERM`s are so prevalent, chance agreement is also very high.)

We should point out that even though we reached a good level of agreement on all of these features, not in all cases it was easy to do so. The only features that are truly easy to annotate are NP type, person, and animacy. Good instructions are needed for gender, number, logical form, multiplicity, deixis, and uniqueness–e.g., for the case of gender one has

to decide what to do with second person pronouns such as *you*, and for deixis the instructions have to specify what to do with objects that are not in the picture although appear to be visible. Finally, the count/mass distinction proved to be very difficult, as did the abstract / concrete distinction (e.g., are diseases abstract or concrete?). We did introduce a number of 'underspecified' values, but this did not lead to results as good as including in the instructions a number of examples (which suggests our scheme may not transport well to other applications).

Antecedent Information

Previous work, particularly in the context of the MUC initiative, suggested that while it's fairly easy to achieve agreement on identity relations, marking up bridging references is quite hard; this was confirmed, e.g., by (Poesio and Vieira, 1998). The only way to achieve a reasonable agreement on this type of annotation, and to contain somehow the annotators' work, is to limit the types of relations annotators are supposed to mark up, and specify priorities. We are currently experimenting with marking up only four types of relations, a subset of those proposed in the 'extended relations' version of the MATE scheme (Poesio et al., 1999a) (which, in turn, derived from Passonneau's DRAMA scheme (Passonneau, 1997): identity (`IDENT`), set membership (`ELEMENT`), subset (`SUBSET`), and 'generalized possession', including part-of relations.

In addition, given our interests we had to be quite strict about the choice of antecedent: whereas in MUC it is perfectly acceptable to mark an 'antecedent' which *follows* a given anaphoric expression, in order, e.g., to compute the CB of an utterance it is necessary to identify the *closest previous* antecedent.

As expected, we are achieving a rather good agreement on identity relations. In our most recent analysis (two annotators looking at the anaphoric relations between 200 NPs) we observed no real disagreements; 79.4% of these relations were marked up by both annotators; 12.8% by only one of them; and in 7.7% of the cases, one of the annotators marked up a closer antecedent than the other. On the other hand, only 22% of bridging references were marked in the same way by both annotators; although our current scheme does limit the disagreements on antecedents and relations (only 4.8% relations are actually marked differently) we still find that 73.17% of relations are marked by only one or the other annotator.

6 EVALUATION

In order to evaluate the compleneteness of our schemes, we have been using the corpus annotated with the reliable features to build statistical models of the process of NP type determination - i.e., the process by which the value of cat is chosen on the basis of the values of the other features. We tried both the Maximum Entropy model (Berger et al., 1996) as implemented by Mikheev (Mikheev, 1998) and the CART model of decision tree construction (Breiman et al., 1984); the results below were obtained using CART. The models are evaluated by comparing the label it predicted on the basis of the features of a given NP with the actual value of cat for that NP, performing a 10-fold cross-validation.

The models discussed in (Poesio et al., 1999b) achieved a 70% accuracy, against a baseline of 22% (if the most common category, `BARE-NP`, is chosen every time.), training on a corpus of 3000 NPs. We are still in the process of evaluating the models built using our second corpus, but partial tests (trained on about 1,000 NPs) suggest that the using the new annotation scheme an accuracy of about 80% can be achieved.

The most complex problem to fix is that of `THIS-NP`s. The reason for the misclassification is that `THIS-NP`s are used in our texts not only to refer to pictures or parts of them, but also to refer to abstract objects introduced by the text, as in the following examples:

(3) a. *A great refinement among armorial signets was to reproduce not only the coat-of-arms but the correct tinctures; they were repeated in colour on the reverse side and the crystal would then be set in the gold bezel. Although the engraved surface could be used for impressions, the colours would not wear away. The signet-ring of Mary, Queen of Scots (beheaded in 1587) is probably the most interesting example of this type;*

b. *The upright secrétaire began to be a fashionable form around the mid-1700s, when letter-writing became a popular past-time. The marchands-merciers were quick to respond to this demand,*

The problem is that such references are difficult to annotate reliably.

7 DISCUSSION

There are some pretty obvious omissions in the work done so far. Even if we only consider the task of NP type determination, there are a number of features whose impact we haven't been able to study so far, in some cases because they proved very hard to annotate. We already discussed two such examples, topichood and thematic roles; another potentially important source of information about the decision to pronominalize, rhetorical structure, is even harder to annotate. We would like to be able to annotate some types of scoping relations as well, especially the cases in which an NP is in the scope of negation as this may license the use of polarity-sensitive items such as *any*. Another important factor is the role of the information which the text planner has decided to realize: e.g., once the text planner has decided to generate both the proper name of discourse entity x, *Alphonse Mucha*, and the fact that x is a Czech painter, the decision to use the `THE-NP` *the Czech painter Alphonse Mucha* is more or less forced on us. And of course, nothing in the scheme discussed above allows us to study the conditions under which a generator may decide to produce a quantifier or a coordinated NPs.

Among the issues raised by this work, an important one is how much of the information that we annotated by hand could be automatically extracted. We believe that a lot of the syntactic information we rely on ($\langle$`unit`$\rangle$ and $\langle$`ne`$\rangle$ identification, $\langle$`unit`$\rangle$ attributes, basic syntactic attributes of $\langle$`ne`$\rangle$) could be extracted automatically using recent advances in robust parsing; this would already cut down the amount of work considerably. The problem is what to do with semantic information: e.g., whether suitable approximations could be found.

Another important question is whether our characterization of NP realization is plausible. One possible objection is that NP type determination goes hand-in-hand with content determination, and the two problems can only be attacked simultaneously. The problem with this type of objection is that it's very difficult to study content determination. This is because of a more general problem with the methodology we are using: there is a mismatch between what a system knows and what an annotator may know about an object–i.e., between the features that a generation system may use and the features that can be annotated, and it's not clear this mismatch can be resolved.

For one thing, the need to choose features that can be annotated reliably imposes serious constraints: features that a generation system can easily set up by itself (e.g., the ILEX system keeps track of what it thinks the current topic is) can be difficult for two annotators to annotate in the same way. Second, some information that a generation system can use when deciding on the type of NP to generate may simply be impossible to annotate. For example, we already seen that the form of an NP often depends on how much information the system intends to communicate to the user about a given entity, or how much information the system believes the user has. In order to build a model of this decision process, we would need to specify for each NP how much information it conveys, and of what type; it's not at all clear that it will be feasible to do this by hand, except in domains in which the annotator knows everything that there is to know about a given object (see, e.g., Jordan's work on the COCONUT domain (Jordan, 1999)).

Conversely, some information that can be annotated - indeed, that is easy to annotate - may not be available to some systems. E.g., we do not know of any system with a lexicon rich enough to specify whether a given entry is functional or not. A solution in this case may be to develop algorithms to extract this information from an annotated corpus, or perhaps just using the syntactic distribution of the predicate as an indication (e.g., a predicate X occurring in a *the X of Y* construction may be functional).

In other words, we believe that the present work is only a first step towards developing an appropriate methodology for empirical investigation and evaluation of generation algorithms, which we nevertheless feel will become more and more necessary. But we believe that, already, this type of work can raise a number of interesting issues concerning semantic annotation and agreement on semantic judgments, which we hope to discuss at the workshop.

Acknowledgments

I wish to thank the other members of the GNOME project, without whom this work would not have been possible– in particular Hua Cheng, Kees van Deemter, Barbara di Eugenio, Renate Henschel, Rodger Kibble, and Rosemary Stevenson. The corpus is being annotated by Debbie De Jongh, Ben

Donaldson, Marisa Flecha-Garcia, Camilla Fraser, Michael Green, Shane Montague, Carol Rennie, and Claire Thomson. I also wish to thank Janet Hitzeman, Pam Jordan, Alistair Knott, Chris Mellish, Johanna Moore, and Jon Oberlander. The GNOME project is supported by the UK Research Council EPSRC, GR/L51126. Massimo Poesio is supported by an EPSRC Advanced Research Fellowship.

References

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72.

L. Breiman, J. H. Friedman, R. A Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Chapman and Hall.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

N. A. Chinchor and B. Sundheim. 1995. Message Understanding Conference (MUC) tests of discourse processing. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–26, Stanford.

S. C. Garrod and A. J. Sanford. 1983. Topic dependent effects in language processing. In G. B. Flores D'Arcais and R. Jarvella, editors, *The Process of Language Comprehension*, pages 271–295. Wiley, Chichester.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225. (The paper originally appeared as an unpublished manuscript in 1986.).

J. K. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

R. Henschel, J. Bateman, and C. Matthiessen. 1999. The solved part of NP generation. In R. Kibble and K. van Deemter, editors, *Proc. of the ESSLLI Workshop on Generating Nominals*, Utrecht.

J. Hitzeman, A. Black, P. Taylor, C. Mellish, and J. Oberlander. 1998. On the use of automatically generated discourse-level information in a concept-to-speech synthesis system. In *Proc. of the International Conference on Spoken Language Processing (ICSLP98),*, page Paper 591, Australia.

P. Jordan. 1999. An empirical study of the communicative goals impacting nominal expressions. In R. Kibble and K. van Deemter, editors, *Proc. of the ESSLLI workshop on The Generation of Nominal Expressions*, Utrecht. University of Utrecht, OTS.

S. Loebner. 1987. Definites. *Journal of Semantics*, 4:279–326.

D. Marcu. 1999. Instructions for manually annotating the discourse structures of texts. Unpublished manuscript, USC/ISI, May.

A. Mikheev. 1998. Feature lattices for maximum entropy modeling. In *Proc. of ACL-COLING*, pages 845–848, Montreal, CA.

J. Oberlander, M. O'Donnell, A. Knott, and C. Mellish. 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–32.

R. Passonneau and D. Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL*.

R. Passonneau. 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript., December.

R. Passonneau. 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, chapter 17, pages 327–358. Oxford University Press.

M. Poesio and R. Stevenson. To appear. *Salience: Computational Models and Psychological Evidence*. Cambridge University Press, Cambridge and New York.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science.

M. Poesio, F. Bruneseaux, and L. Romary. 1999a. The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker, editor, *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

M. Poesio, R. Henschel, J. Hitzeman, R. Kibble, S. Montague, and K. van Deemter. 1999b. Towards an annotation scheme for Noun Phrase generation. In B. Krenn H. Uszkoreit, T. Brants, editor, *Proc. of the EACL workshop on Linguistically Interpreted Corpora (LINC-99)*.

E. F. Prince. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

T. Reinhart. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1). Also distributed by Indiana University Linguistics Club.

D. Scott, R. Power, and R. Evans. 1998. Generation as a solution to its own problem. In *Proc. of the 9th International Workshop on Natural Language Generation*, Niagara-on-the-Lake, CA.