REFERENTIAL SEMANTIC LANGUAGE MODELING FOR DATA-POOR DOMAINS

Stephen Wu, Lane Schwartz, and William Schuler

University of Minnesota
Department of Computer Science and Engineering
Minneapolis, MN
{swu,lschwar,schuler}@cs.umn.edu

ABSTRACT

This paper describes a referential semantic language model that achieves accurate recognition in user-defined domains with no available domain-specific training corpora. This model is interesting in that, unlike similar recent systems, it exploits context *dynamically*, using incremental processing and limited stack memory of an HMM-like time series model to constrain search.

***Index Terms*—** Natural language interfaces, Speech recognition, Artificial intelligence

1. INTRODUCTION

The development of general-purpose artificial assistants could have a transformative effect on society from early education to elder care. But to be useful, these assistants will need to communicate with the people they assist in the mutable and idiosyncratic language of day to day life, populated with proper names of co-workers, objects, and local events not found in broad corpora. Language models generally require training corpora of example sentences, but interfaces for assistants can exploit another source of information: a model of the world with which they are expected to assist. This is an extremely valuable resource – if the world model is mostly known by the user, or even created by the user through the interface, hypothesized directives that do indeed describe entities in the world model are much more likely to be correct than those that do not.

This paper describes a framework for incorporating referential semantic information from a world model directly into a probabilistic language model, rather than relying solely on phonologic and syntactic information. Introducing world model referents into the decoding search greatly increases the search space, but the decoder can incrementally prune this search based on probabilities associated with combined phonological, syntactic, and referential contexts.

This model is incremental in that interpretation is performed in the order in which words are received. However, unlike earlier constraint-based incremental interpreters [1, 2], the approach described in this paper pursues multiple interpretations at once, ranked probabilistically. Moreover, unlike more recent speech recognizers which constrain search based on pre-compiled word n-grams [3, 4], this approach can be applied in mutable environments without expensive pre-compilation, and can exploit *intra-sentential* contexts[1]. Finally, since this approach performs interpretation based on the left-to-right sharing of a Viterbi dynamic programming algorithm instead of the bottom-up sharing of a CKY-like parsing algorithm [5, 6, 7, 8], *inter-sentential* context can constrain semantics at the beginning of recognition, avoiding the relatively unconstrained sets of referents which arise at the bottom of a parser chart.

2. BACKGROUND

2.1. Referential Semantics

The language model described in this paper defines semantic referents in terms of a world model $\mathcal{M}$. In model theory [9, 10], a world model is defined as a tuple $\mathcal{M} = \langle \mathcal{E}, [\![\cdot]\!] \rangle$ containing a domain of entity constants $\mathcal{E}$ and an interpretation function $[\![\cdot]\!]$ to interpret expressions in terms of those constants. Here, $[\![\cdot]\!]$ is quite versatile, accepting expressions ϕ that are logical statements (simple type $\mathbf{T}$), references to entities (simple type $\mathbf{E}$), or functors (complex type $\langle \alpha, \beta \rangle$) that take an argument of type α and produce output of type β. These functor expressions ϕ can then be applied to other expressions ψ of type α as arguments to yield expressions $\phi(\psi)$ of type β. By nesting functors, complex expressions can be defined, denoting sets or properties of entities: $\langle \mathbf{E}, \mathbf{T} \rangle$, relations over entity pairs: $\langle \mathbf{E}, \langle \mathbf{E}, \mathbf{T} \rangle \rangle$, or higher-order functors over sets: $\langle \langle \mathbf{E}, \mathbf{T} \rangle, \langle \mathbf{E}, \mathbf{T} \rangle \rangle$.

First order or higher models (in which functors can take sets as arguments) can be mapped to equivalent zero order models (with functors defined only on entities). This is generally motivated by a desire to allow sets of entities to be described in much the same way as individual entities [11]. Entities in a zero order model $\mathcal{M}$ can be defined from entities in a higher order model $\mathcal{M}'$ by mapping (or *reifying*) each set $S = \{\mathbf{e}'_1, \mathbf{e}'_2, \dots\}$ in $\mathcal{P}(\mathcal{E}_{\mathcal{M}'})$ (or set of sets in $\mathcal{P}(\mathcal{P}(\mathcal{E}_{\mathcal{M}'}))$, etc.) as an entity $\mathbf{e_S}$ in $\mathcal{E}_{\mathcal{M}}$.[2] Zero order functors in the interpretation function of $\mathcal{M}$ can be defined directly from higher order functors (over sets) in $\mathcal{M}'$ by mapping each instance of $\langle S_1, S_2 \rangle$ in $[\![l']\!]_{\mathcal{M}'} : \mathcal{P}(\mathcal{E}_{\mathcal{M}'}) \times \mathcal{P}(\mathcal{E}_{\mathcal{M}'})$ to a corresponding instance of $\langle \mathbf{e_{S_1}}, \mathbf{e_{S_2}} \rangle$ in $[\![l]\!]_{\mathcal{M}} : \mathcal{E}_{\mathcal{M}} \times \mathcal{E}_{\mathcal{M}}$. Set subsumption $\mathcal{M}'$ can then be defined on entities made from reified sets in $\mathcal{M}$, similar to 'ISA' relations over concepts in knowledge representation systems [12]. These relations can be represented in a lattice, as shown in Figure 1.

2.2. Language Modeling and Hierarchic HMMs

The model described in this paper is a specialization of the Hidden Markov Model (HMM) framework commonly used in speech recognition [13, 14]. HMMs characterize speech as a sequence of hidden states q_t (which may consist of speech sounds, words, or other hypothesized syntactic or semantic information), and observed states a_t (typically short, overlapping frames of an audio signal) at

This research was supported by NSF CAREER award 0447685. The views expressed are not necessarily endorsed by the sponsors.

[1] For example, the initial semantic context of 'go to the garage workbench, and get ... ' gives a powerful constraint on possible completions.

[2] Here, $\mathcal{P}(X)$ is the power set of X, containing the set of all subsets.

corresponding time steps t. A most probable sequence of hidden states $\hat{q}_{1..T}$ can then be hypothesized given any sequence of observed states $o_{1..T}$, using Bayes' Law (Equation 2) and Markovian independence assumptions (Equation 3) to define the full $\mathsf{P}(q_{1..T} \mid a_{1..T})$ probability as the product of a *Language Model (LM)* prior probability $\mathsf{P}(q_{1..T}) \overset{\text{def}}{=} \prod_t \mathsf{P}_{\Theta_{\text{LM}}}(q_t \mid q_{t-1})$ and an *Acoustical Model (AM)* likelihood probability $\mathsf{P}(a_{1..T} \mid q_{1..T}) \overset{\text{def}}{=} \prod_t \mathsf{P}_{\Theta_{\text{AM}}}(a_t \mid q_t)$:

$$\hat{q}_{1..T} = \underset{q_{1..T}}{\text{argmax}}\, \mathsf{P}(q_{1..T} \mid a_{1..T}) \tag{1}$$

$$= \underset{q_{1..T}}{\text{argmax}} \mathsf{P}(q_{1..T}) \cdot \mathsf{P}(a_{1..T} \mid q_{1..T}) \tag{2}$$

$$\overset{\text{def}}{=} \underset{q_{1..T}}{\text{argmax}} \prod_{t=1}^{T} \mathsf{P}_{\Theta_{\text{LM}}}(q_t \mid q_{t-1}) \cdot \mathsf{P}_{\Theta_{\text{AM}}}(a_t \mid q_t) \tag{3}$$

Hierarchic Hidden Markov Models (HHMMs) [15] model language model transitions $\mathsf{P}(\alpha_t \mid \alpha_{t-1})$ using hierarchies of component HMMs. Overall, transition probabilities are calculated in two phases: a 'reduce' phase (resulting in an intermediate state β), in which component HMMs may terminate; and a 'shift' phase (resulting in a modeled state α_t), in which unterminated HMMs transition, and terminated HMMs are re-initialized from their parent HMMs. Variables over intermediate and modeled states are factored into sequences of depth-specific variables – one for each of the D levels in the HMM hierarchy:

$$\alpha_t = \langle \alpha_t^1 \ldots \alpha_t^D \rangle \tag{4}$$

$$\beta = \langle \beta^1 \ldots \beta^D \rangle \tag{5}$$

Transition probabilities are then calculated as a product of transition probabilities at each level:

$$\mathsf{P}(\alpha_t \mid \alpha_{t-1}) = \sum_{\beta} \mathsf{P}(\beta \mid \alpha_{t-1}) \cdot \mathsf{P}(\alpha_t \mid \beta\, \alpha_{t-1}) \tag{6}$$

$$\overset{\text{def}}{=} \sum_{\beta^1 \ldots \beta^D} \left[\prod_{d=1}^{D} \mathsf{P}_{\Theta_\beta}(\beta^d \mid \beta^{d+1} \alpha_{t-1}^d) \right] \cdot \left[\prod_{d=1}^{D} \mathsf{P}_{\Theta_\alpha}(\alpha_t^d \mid \beta^d \beta^{d+1} \alpha_t^{d-1} \alpha_{t-1}^d) \right] \tag{7}$$

with $\beta^{D+1} = \beta_\perp$ and $\alpha_t^0 = \alpha_\top$.

In Murphy-Paskin HHMMs, each modeled state variable α_t^d is a syntactic, lexical, or phonetic category q_t^d and each intermediate state variable β^d is a boolean switching variable $f^d \in \{\mathbf{0}, \mathbf{1}\}$.

$$\alpha_t^d = q_t^d \tag{8}$$

$$\beta^d = f^d \tag{9}$$

Instantiating Θ_β as $\Theta_{\text{MP-}\beta}$, f^d is true when there is a transition at the level below d and the stack element q_{t-1}^d is a final state:[3]

$$\mathsf{P}_{\Theta_{\text{MP-}\beta}}(f^d \mid f^{d+1} q_{t-1}^d) \overset{\text{def}}{=} \begin{cases} \text{if } f^{d+1}{=}0 & : |f^d{=}0| \\ \text{if } f^{d+1}{=}1,\ q_{t-1}^d \in \mathit{Final} & : |f^d{=}1| \\ \text{if } f^{d+1}{=}1,\ q_{t-1}^d \notin \mathit{Final} & : |f^d{=}0| \end{cases} \tag{10}$$

and shift probabilities at each level (instantiating Θ_α as $\Theta_{\text{MP-}\alpha}$) are:

$$\mathsf{P}_{\Theta_{\text{MP-}\alpha}}(q_t^d \mid f^d f^{d+1} q_t^{d-1} q_{t-1}^d) \overset{\text{def}}{=} \begin{cases} \text{if } f^d{=}0,\ f^{d+1}{=}0 : |q_t^d{=}q_{t-1}^d| \\ \text{if } f^d{=}0,\ f^{d+1}{=}1 : \mathsf{P}_{\Theta_{\text{MP-Trans}}}(q_t^d \mid q_{t-1}^d) \\ \text{if } f^d{=}1,\ f^{d+1}{=}1 : \mathsf{P}_{\Theta_{\text{MP-Init}}}(q_t^d \mid q_t^{d-1}) \end{cases} \tag{11}$$

and $\mathbf{f}_\perp = \mathbf{1}$ and $\mathbf{q}_\top = \mathbf{ROOT}$.

3. REFERENTIAL SEMANTIC DECODING

A referential semantic language model can now be defined as an instantiation of an HHMM, interpreting directives in a reified world model.

3.1. Dynamic Reference

The model of reference described in this paper is interesting in that it *transitions* through time, basing the value of e at each time t on the previous value of e, at time $t{-}1$. When a word w and associated relation l is hypothesized, a referent e transitions from e_{t-1} to e_t, where e_{t-1} is a hypothesized referent described by the utterance prior to w, and e_t is the result of additionally constraining e_{t-1} by $[\![l]\!]_{\mathcal{M}}$ the meaning of w in $\mathcal{M}$. The model is dynamically context sensitive in the sense that referents e_t are defined in context of referent e_{t-1}.

The language model interacts with $\mathcal{M}$ through queries of the form $[\![l]\!]_{\mathcal{M}}(\mathbf{e}_{\mathbf{S_1}}, \mathbf{e}_{\mathbf{S_2}})$, where $\mathbf{e}_{\mathbf{S_1}}$ is an argument referent (if l is a relation), and $\mathbf{e}_{\mathbf{S_2}}$ is a context referent. Recall the definition in Section 2.1 of a zero-order model $\mathcal{M}$ with entities $\mathbf{e}_{\{\mathbf{e}', \mathbf{e}'', \ldots\}}$ reified from sets of individuals $\{\mathbf{e}', \mathbf{e}'', \ldots\}$ in a first- or higher-order model $\mathcal{M}'$. The context-sensitive reference model described in this paper is conditioned on relations l over the reified sets in $\mathcal{M}$, which are defined in terms of corresponding relations l' in $\mathcal{M}'$:

if $[\![l']\!]_{\mathcal{M}'}$ is of type $\langle \mathbf{E}, \mathbf{T} \rangle$:

$$[\![l]\!]_{\mathcal{M}}(\mathbf{e}_{\mathbf{S_1}}, \mathbf{e}_{\mathbf{S_2}}) = \mathbf{e}_{\mathbf{S}} \text{ iff } S = S_2 \cap [\![l']\!]_{\mathcal{M}'} \tag{12}$$

if $[\![l']\!]_{\mathcal{M}'}$ is of type $\langle \mathbf{E}, \langle \mathbf{E}, \mathbf{T} \rangle \rangle$:

$$[\![l]\!]_{\mathcal{M}}(\mathbf{e}_{\mathbf{S_1}}, \mathbf{e}_{\mathbf{S_2}}) = \mathbf{e}_{\mathbf{S}} \text{ iff } S = S_2 \cap (S_1 \cdot [\![l']\!]_{\mathcal{M}'}) \tag{13}$$

if $[\![l']\!]_{\mathcal{M}'}$ is of type $\langle \langle \mathbf{E}, \mathbf{T} \rangle, \langle \mathbf{E}, \mathbf{T} \rangle \rangle$:

$$[\![l]\!]_{\mathcal{M}}(\mathbf{e}_{\mathbf{S_1}}, \mathbf{e}_{\mathbf{S_2}}) = \mathbf{e}_{\mathbf{S}} \text{ iff } S = S_2 \cap [\![l']\!]_{\mathcal{M}'}(S_1) \tag{14}$$

where relation products are defined to resemble matrix products:

$$S \cdot R = \{e'' \mid e' \in S,\ \langle e', e'' \rangle \in R\} \tag{15}$$

Note that in each case above, the set of referents S corresponding to the reified output of $[\![l]\!]_{\mathcal{M}}$ results from an intersection of the set S_2 corresponding to the last argument of $[\![l]\!]_{\mathcal{M}}$. This set S_2 is the context. All intersections with this context referent result in a transition from a less constrained referent (corresponding to a larger set in $\mathcal{M}'$) to a more constrained referent, (corresponding to a smaller set in $\mathcal{M}'$). These intersections can be viewed on a subsumption lattice (see Figure 1).

3.2. A Referential Semantic Language Model

The referential semantic language model decomposes the HHMM stack variables α_t^d at each depth d and time step t into semantic referent (reified entity) e_t^d and syntactic category c_t^d variables; and decomposes the HHMM reduce variables β^d into reduced referent e_R^d and final state f_R^d variables:

$$\alpha_t^d = \langle e_t^d, c_t^d \rangle \tag{16}$$

$$\beta^d = \langle e_R^d, f_R^d \rangle \tag{17}$$

[3] Here $|\cdot|$ is an indicator function: $|\phi| = 1$ if ϕ is true, 0 otherwise.

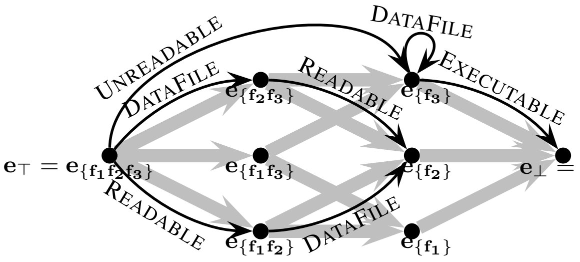
UNREADABLE
DATAFILE
DATAFILE
EXECUTABLE
READABLE
READABLE
DATAFILE
$\mathbf{e}_\top = \mathbf{e}_{\{\mathbf{f_1 f_2 f_3}\}}$
$\mathbf{e}_{\{\mathbf{f_2 f_3}\}}$
$\mathbf{e}_{\{\mathbf{f_1 f_3}\}}$
$\mathbf{e}_{\{\mathbf{f_1 f_2}\}}$
$\mathbf{e}_{\{\mathbf{f_3}\}}$
$\mathbf{e}_{\{\mathbf{f_2}\}}$
$\mathbf{e}_{\{\mathbf{f_1}\}}$
$\mathbf{e}_\bot = \mathbf{e}_\emptyset$

Fig. 1. A subsumption lattice (laid on its side, in gray) over the power set of a domain containing three files: f_1 (a readable executable), f_2 (a readable data file), and f_3 (an unreadable data file). 'Reference paths' made up of conjunctions of relations l (directed arcs, in black) traverse the lattice from left to right toward the empty set, as referents ($\mathbf{e}_{\{\dots\}}$, corresponding to sets of files) are incrementally constrained by intersection with each $[\![l]\!]_{\mathcal{M}}$. (Some arcs are omitted for clarity.)

Reduce probabilities at each level (instantiating Θ_β as $\Theta_{\text{RSLM-}\beta}$) are:

$$\begin{aligned}\mathsf{P}_{\Theta_{\text{RSLM-}\beta}}(\langle e^d_R\ f^d_R\rangle \mid \langle e^{d+1}_R f^{d+1}_R\rangle\langle e^d_{t-1} c^d_{t-1}\rangle) \\ \stackrel{\text{def}}{=} |e^d_R = [\![\textit{label-end}(c^d_t)]\!]_{\mathcal{M}}(e^{d+1}_R, e^d_{t-1})| \\ \cdot\, \mathsf{P}_{\Theta_{\text{MP-}\beta}}(f^d_R \mid f^{d+1}_R c^d_{t-1})\end{aligned} \tag{18}$$

where *label-end*(c^d_t) defines a functor in $[\![\cdot]\!]_{\mathcal{M}}$ at HMM final state c^d_t to compose the result of the HMM at depth d with that at depth $d{-}1$. Shift probabilities at each level (instantiating Θ_α as $\Theta_{\text{RSLM-}\alpha}$) are:

$$\begin{aligned}&\mathsf{P}_{\Theta_{\text{RSLM-}\alpha}}(\langle e^d_t c^d_t\rangle \mid \langle e^d_R\ f^d_R\rangle\langle e^{d+1}_R f^{d+1}_R\rangle\langle e^{d-1}_t c^{d-1}_t\rangle\langle e^d_{t-1} c^d_{t-1}\rangle) \\ &\stackrel{\text{def}}{=} \begin{cases}\text{if } f^d_R{=}0,\ f^{d+1}_R{=}0 : |e^d_t{=}e^d_R| \cdot |c^d_t{=}c^d_{t-1}| \\ \text{if } f^d_R{=}0,\ f^{d+1}_R{=}1 : |e^d_t{=}e^d_R| \cdot \mathsf{P}_{\Theta_{\text{Syn-Trans}}}(c^d_t \mid c^d_{t-1}) \\ \text{if } f^d_R{=}1,\ f^{d+1}_R{=}1 : \sum_l \mathsf{P}_{\Theta_{\text{Ref-Init}}}(l \mid e^{d-1}_t c^{d-1}_t) \\ \qquad\qquad \cdot|e^d_t{=}[\![l]\!]_{\mathcal{M}}(e^{d-1}_t, \mathbf{e}_\top)| \\ \qquad\qquad \cdot\mathsf{P}_{\Theta_{\text{Syn-Init}}}(c^d_t \mid l\ c^{d-1}_t)\end{cases}\end{aligned} \tag{19}$$

and $r^{D+1} = \langle e^D_{t-1}, \mathbf{1}\rangle$ and $s^0_t = \langle \mathbf{e}_\top, \mathbf{ROOT}\rangle$.

3.3. Reference Transitions on a Subsumption Lattice

This model treats properties (unary relations like READABLE or DATAFILE) as labeled transitions l' on a subsumption lattice from supersets e_{t-1} to subsets e_t that result from intersecting e_{t-1} with $[\![l']\!]_{\mathcal{M}'}$ (see Figure 1).[4]

A general template for intersective adjectives can be expressed as a noun phrase (NP) expansion using the following regular expression:

$$\text{NP}(g) \rightarrow \text{Det}\ \left(\ \text{Adj}(g)\ \right)^*\ \text{Noun:}l(g)\ \left(\ \text{PP}(g) \mid \text{RC}(g)\ \right)^*$$

where g is a variable over referential contexts (in this case, reified sets of individuals that are considered potential referents while the noun phrase is being interpreted), which is successively constrained by the semantics of the adjective and noun relation l, followed by optional prepositional phrase (PP) and relative clause (RC) modifiers.

3.4. Reference Transitions with Relation Arguments

Sequences of properties (unary relations) can be interpreted as simple nonbranching paths in a subsumption lattice, but higher-arity relations define more complex paths that fork and rejoin. As an example of PP or RC modifiers, the set of directories(set g) that '*contain* things that are *user-readable objects*' would be reachable only by:

1. pushing the original set of directories g onto a referent stack,
2. traversing a CONTAIN relation departing g to obtain the contents of those directories h,
3. traversing a READABLE relation departing h to constrain this set to the set of contents that are also user-readable objects,
4. traversing the inverse CONTAIN^I of relation CONTAIN to obtain the containers of these user-readable objects, then constraining the original set of directories g by intersection with this resulting set to yield the directories containing user-readable objects.

'Forking' is therefore handled via syntactic recursion: one path is explored by the recognizer while the other waits on a stack. A general template for branching reduced relative clauses (or prepositional phrases) that exhibit this forking behavior can be expressed as below, using the variables g and h defined above:

$$\text{RC}(g) \rightarrow \text{Verb:}l(g,h)\ \ \text{NP}(h)\ \ -\!:\!l^I(h,g)$$

where the inverse or transpose relation l^I at the last, empty constituent '$-$' is intended to apply when the NP expansion concludes or reduces (this relation l^I is returned by the *end-label* function described earlier).

3.5. Training

Although linguistic training data for the envisaged applications of this model are likely to be scarce, the reference model ($\Theta_{\text{Ref-Init}}$) introduced in Equation 19 can in principle be trained on non-linguistic examples of how the interfaced system is used (e.g. which referents in a world model are more likely to be modified). In the evaluation described below, however, these were all set to uniform distributions over the arcs departing each context referent.

The syntactic models $\Theta_{\text{Syn-Init}}$ and $\Theta_{\text{Syn-Trans}}$ in Equation 19 can in principle be trained on-line, assuming non-zero priors for new words. Again, however, in the evaluation described below, these were all set to uniform over all regular expressions matching each appropriate context.

Models not described in this paper, including pronunciation, subphone transition, and acoustical models, were either taken directly from the Robinson RNN recognizer [16], or were provided in the same way as described there (e.g. from a pronunciation lexicon).

4. EVALUATION

To evaluate the contribution to recognition accuracy of referential semantics over that of syntax and phonology alone, a baseline (syntax only) and test (baseline plus referential semantics) recognizer were run on sample ontology manipulation directives in a benchmark 'student activities' domain.

4.1. A Student Activities Database

The student activities ontology organizes extracurricular activities under subcategories (e.g. sports ⊃ football ⊃ offense), and organizes

[4] This lattice need not be an actual data structure. Since the world model is queried incrementally, the lattice relations may be calculated as needed.

students into homerooms, in which context they can be identified by a first or last name. Every student or activity is an entity e in the set of entities $\mathcal{E}$, and relations l are subcategories' or persons' names.

The original student activities world model $\mathcal{M}_{240}$ includes 240 entities in $\mathcal{E}$: 158 categories (groups or positions) and 82 instances (students), each connected via a labeled arc from a parent category. An expanded version of the students ontology, $\mathcal{M}_{4175}$, includes 4175 entities from 717 concepts and 3458 instances. The extra entities are merely distractors to the referents in $\mathcal{M}_{240}$, which remain intact.

This ontology is manipulated using directives such as:

(1) 'set homeroom two, Bell, to sports, football, captain'

which are incrementally interpreted by transitioning down the subsumption lattice (e.g. from 'sports' to 'football' to 'captain') or forking to another part of the lattice (e.g. from 'Bell' to 'sports').

4.2. Empirical Results

A corpus of 144 test sentences (no training sentences) was collected from 7 native English speakers (5 male, 2 female), who were asked to make specific edits to the student activities ontology described above.[5] The average sentence length in this collection is 7.17 words.

Baseline and test versions of this system were run using a RNN acoustical model [16] trained on the TIMIT corpus of read speech [17]. Results below report concept error rate (CER), where concepts correspond to relation labels in the world model.[6]

test	correct	subst	delete	insert	CER
$\mathcal{M}_{240}$	86.4	11.3	2.34	3.41	17.1
$\mathcal{M}_{4175}$	84.5	13.5	2.05	4.39	19.9
$\mathcal{M}_0$	67.1	27.5	5.46	10.5	43.5
trigram from $\mathcal{M}_{240}$	78.1	15.0	6.92	4.68	26.6

Results using the initial world model with 240 entities ($\mathcal{M}_{240}$) show an overall 17.1% concept error rate. These directives, tested with additional distracting referents in $\mathcal{M}_{4175}$, shows a slight CER increase to 19.9%. The use of this world model with no linguistic training data is comparable to that reported for other systems, which *were* trained on sample sentences [4, 3].

In comparison, a baseline using only the grammar from the students domain without any world model information and no linguistic training data ($\mathcal{M}_0$) scores a CER of 43.5%, which is significantly higher ($p = 1.1 \times 10^{-19}$ using pairwise t-test with $\mathcal{M}_{240}$). A 'compromise' word trigram language model compiled from the referential semantic model above (in the 240-entity domain) scores 26.6%, also significantly higher error than $\mathcal{M}_{240}$ ($p = 3.2 \times 10^{-5}$ using pairwise t-test), suggesting that referential context is more predictive than n-gram context. Moreover, this compilation to trigrams is impractically expensive (requiring several hours of pre-processing), as it must consider all combinations of entities in the world model.[7]

Though referents neglected by the beam early in the utterance can cause recognition errors later, a sufficiently large beam mitigates this effect. All evaluations ran in real time with a beam width of 1000 hypotheses per frame on an 8-processor 2.6GHz server.

5. CONCLUSION

This paper has described a language model that achieves accurate recognition in user-defined domains with no available domain-specific training corpora, through the use of explicit hypothesized semantic referents. This architecture requires that the interfaced application make available a queriable world model, but the combined phonological, syntactic, and referential semantic decoding process ensures the world model is only queried when necessary, allowing accurate real time performance even in large domains containing several thousand entities.

6. REFERENCES

[1] Nicholas Haddock, "Computational models of incremental semantic interpretation," *Language and Cognitive Processes*, vol. 4, pp. 337–368, 1989.

[2] Chris Mellish, *Computer interpretation of natural language descriptions*, Wiley, New York, 1985.

[3] Oliver Lemon and Alexander Gruenstein, "Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments," *ACM Transactions on Computer-Human Interaction*, vol. 11, no. 3, pp. 241–267, 2004.

[4] G. Chung, S. Seneff, C. Wang, and I. Hetherington, "A dynamic vocabulary spoken dialogue interface," in *Proc. ICSLP*, 2004, pp. 1457–1460.

[5] William Schuler, "Computational properties of environment-based disambiguation," in *Proc. ACL*, 2001, pp. 466–473.

[6] David DeVault and Matthew Stone, "Domain inference in incremental interpretation," in *Proc. ICoS*, 2003, pp. 73–87.

[7] Peter Gorniak and Deb Roy, "Grounded semantic composition for visual scenes," *Journal of Artificial Intelligence Research*, vol. 21, pp. 429–470, 2004.

[8] Gregory Aist, James Allen, Ellen Campana, Carlos Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus, "Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods," in *Proc. DECALOG*, 2007, pp. 149–154.

[9] Alfred Tarski, "The concept of truth in the languages of the deductive sciences (polish)," *Prace Towarzystwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych*, vol. 34, 1933, translated as 'The concept of truth in formalized languages', in: J. Corcoran (Ed.), Logic, Semantics, Metamathematics: papers from 1923 to 1938, Hackett Publishing Company, Indianapolis, IN, 1983, pp. 152–278.

[10] Alonzo Church, "A formulation of the simple theory of types," *Journal of Symbolic Logic*, vol. 5, no. 2, pp. 56–68, 1940.

[11] Jerry R. Hobbs, "Ontological promiscuity," in *Proc. ACL*, 1985, pp. 61–69.

[12] Ronald J. Brachman and James G. Schmolze, "An overview of the kl-one knolewdge representation system," *Cognitive Science*, vol. 9, no. 2, pp. 171–216, Apr. 1985.

[13] James Baker, "The Dragon system: an overivew," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

[5] References to entities not found in the world model can be recognized, but should be dispreferred in most applications.

[6] In this domain, directives are mostly sequences of relation labels, so nearly every word is a concept.

[7] As time-series models, HHMMs can also be directly interporlated with word n-gram models – but an analysis of the resulting model is beyond the scope of this paper.

[14] Frederick Jelinek, Lalit R. Bahl, and Robert L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Transactions on Information Theory*, vol. 21, pp. 250–256, 1975.

[15] Kevin P. Murphy and Mark A. Paskin, "Linear time inference in hierarchical HMMs," in *Proc. NIPS*, 2001, pp. 833–840.

[16] Tony Robinson, "An application of recurrent nets to phone probability estimation," in *IEEE Transactions on Neural Networks*, 1994, vol. 5, pp. 298–305.

[17] William M. Fisher, Victor Zue, Jared Bernstein, and David S. Pallet, "An acoustic-phonetic data base," *Journal of the Acoustical Society of America*, vol. 81, pp. S92–S93, 1987.