A Probabilistic Rich Type Theory for Semantic Interpretation

Robin Cooper[1], **Simon Dobnik**[1], **Shalom Lappin**[2], **and Staffan Larsson**[1]
[1]University of Gothenburg, [2]King's College London
{cooper,sl}@ling.gu.se, simon.dobnik@gu.se, shalom.lappin@kcl.ac.uk

Abstract

We propose a probabilistic type theory in which a situation s is judged to be of a type T with probability p. In addition to basic and functional types it includes, *inter alia*, record types and a notion of typing based on them. The type system is intensional in that types of situations are not reduced to sets of situations. We specify the fragment of a compositional semantics in which truth conditions are replaced by probability conditions. The type system is the interface between classifying situations in perception and computing the semantic interpretations of phrases in natural language.

1 Introduction

Classical semantic theories (Montague, 1974), as well as dynamic (Kamp and Reyle, 1993) and underspecified (Fox and Lappin, 2010) frameworks use categorical type systems. A type T identifies a set of possible denotations for expressions in T, and the system specifies combinatorial operations for deriving the denotation of an expression from the values of its constituents.

These theories cannot represent the gradience of semantic properties that is pervasive in speakers' judgements concerning truth, predication, and meaning relations. In general, predicates do not have determinate extensions (or intensions), and so, in many cases, speakers do not make categorical judgements about the interpretation of an expression. Attributing gradience effects to performance mechanisms offers no help, unless one can show precisely how these mechanisms produce the observed effects.

Moreover, there is a fair amount of evidence indicating that language acquisition in general crucially relies on probabilistic learning (Clark and Lappin, 2011). It is not clear how a reasonable account of semantic learning could be constructed on the basis of the categorical type systems that either classical or revised semantic theories assume. Such systems do not appear to be efficiently learnable from the primary linguistic data (with weak learning biases), nor is there much psychological data to suggest that they provide biologically determined constraints on semantic learning.

A semantic theory that assigns probability rather than truth conditions to sentences is in a better position to deal with both of these issues. Gradience is intrinsic to the theory by virtue of the fact that speakers assign values to declarative sentences in the continuum of real numbers [0,1], rather than Boolean values in $\{0,1\}$. In addition, a probabilistic account of semantic learning is facilitated if the target of learning is a probabilistic representation of meaning. Both semantic representation and learning are instances of reasoning under uncertainty.

Probability theorists working in AI often describe probability judgements as involving distributions over worlds. In fact, they tend to limit such judgements to a restricted set of outcomes or events, each of which corresponds to a partial world which is, effectively, a type of situation (Halpern, 2003). A classic example of the reduction of worlds to situation types in probability theory is the estimation of the likelihood of heads vs tails in a series of coin tosses. Here the world is held constant except along the dimension of a binary choice between a particular set of possible outcomes. A slightly more complex case is the probability distribution for possible results of throwing a single die, which allows for six possibilities corresponding to each of its numbered faces. This restricted range of outcomes constitutes the sample space.

We are making explicit the assumption, common to most probability theories used in AI, with clearly defined sample spaces, that probability is distributed over situation types (Barwise and Perry, 1983), rather than over sets of entire worlds. An Austinian proposition is a judgement that a

Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS), pages 72–79,
Gothenburg, Sweden, April 26-30 2014.

situation is of a particular type, and we treat it as probabilistic. In fact, it expresses a subjective probability in that it encodes the belief of an agent concerning the likelihood that a situation is of that type. The core of an Austinian proposition is a type judgement of the form $s : T$, which states that a situation s is of type T. On our account this judgement is expressed probabilistically as $p(s : T) = r$, where $r \in [0,1]$.[1]

On the probabilistic type system that we propose situation types are intensional objects over which probability distributions are specified. This allows us to reason about the likelihood of alternative states of affairs without invoking possible worlds.

Complete worlds are not tractably representable. Assume that worlds are maximal consistent sets of propositions (Carnap, 1947). If the logic of propositions is higher-order, then the problem of determining membership in such a set is not complete. If the logic is classically first-order, then the membership problem is complete, but undecidable.

Alternatively, we could limit ourselves to propositional logic, and try to generate a maximally consistent set of propositions from a single finite proposition P in Conjunctive Normal Form (CNF, a conjunction of disjunctions), by simply adding conjuncts to P. But it is not clear what (finite) set of rules or procedures we could use to decide which propositions to add in order to generate a full description of a world in a systematic way. Nor is it obvious at what point the conjunction will constitute a complete description of the world.

Moreover, all the propositions that P entails must be added to it, and all the propositions with which P is inconsistent must be excluded, in order to obtain the maximal consistent set of propositions that describe a world. But then testing the satisfiability of P is an instance of the $ksat$ problem, which, in the general case, is NP-complete.[2]

[1] Beltagy et al. (2013) propose an approach on which classical logic-based representations are combined with distributional lexical semantics and a probabilistic Markov logic, in order to select among the set of possible inferences from a sentence. Our concern here is more foundational. We seek to replace classical semantic representations with a rich probabilistic type theory as the basis of both lexical and compositional interpretation.

[2] The $ksat$ problem is to determine whether a formula in propositional logic has a satisfying set of truth-value assignments. For the complexity results of different types of $ksat$ problem see Papadimitriou (1995).

By contrast situation types can be as large or as small as we need them to be. They are not maximal in the way that worlds are, and so the issue of completeness of specification does not arise. Therefore, they can, in principle, be tractably represented.

2 Rich Type Theory and Probability

Central to standard formulations of rich type theories (for example, (Martin-Löf, 1984)) is the notion of a judgement $a : T$, that object a is of type T. We represent the probability of this judgement as $p(a : T)$. Our system (based on Cooper (2012)) includes the following types.

Basic Types are not constructed out of other objects introduced in the theory. If T is a basic type, $p(a : T)$ for any object a is provided by a probability model, an assignment of probabilities to judgements involving basic types.

PTypes are constructed from a *predicate* and an appropriate sequence of arguments. An example is the predicate 'man' with arity $\langle Ind, Time\rangle$ where the types *Ind* and *Time* are the basic type of individuals and of time points respectively. Thus *man(john,18:10)* is the type of situation (or eventuality) where John is a man at time 18:10. A probability model provides probabilities $p(e : r(a_1, \ldots, a_n))$ for ptypes $r(a_1, \ldots, a_n)$. We take both common nouns and verbs to provide the components out of which PTypes are constructed.

Meets and Joins give, for T_1 and T_2, the meet, $T_1 \wedge T_2$ and the join $T_1 \vee T_2$, respectively. $a : T_1 \wedge T_2$ just in case $a : T_1$ and $a : T_2$. $a : T_1 \vee T_2$ just in case either $a : T_1$ or $a : T_2$ (possibly both).[3] The probabilities for meet and joint types are defined by the classical (Kolmogorov, 1950) equations $p(a : T_1 \wedge T_2) = p(a : T_1)p(a : T_2 \mid a : T_1)$ (equivalently, $p(a : T_1 \wedge T_2) = p(a : T_1, a : T_2)$), and $p(a : T_1 \vee T_2) = p(a : T_1) + p(a : T_2) - p(a : T_1 \wedge T_2)$, respectively.

Subtypes A type T_1 is a subtype of type T_2, $T_1 \sqsubseteq T_2$, just in case $a : T_1$ implies $a : T_2$ no matter what we assign to the basic types. If $T_1 \sqsubseteq T_2$ then $a : T_1 \wedge T_2$ iff $a : T_1$ and $a : T_1 \vee T_2$ iff $a : T_2$. Similarly, if $T_2 \sqsubseteq T_1$ then $a : T_1 \wedge T_2$ iff $a : T_2$ and $a : T_1 \vee T_2$ iff $a : T_1$.

If $T_2 \sqsubseteq T_1$, then $p(a : T_1 \wedge T_2) = p(a : T_2)$, and $p(a : T_1 \vee T_2) = p(a : T_1)$. If $T_1 \sqsubseteq T_2$,

[3] This use of intersection and union types is not standard in rich type theories, where product and disjoint union are preferred following the Curry-Howard correspondence for conjunction and disjunction.

then $p(a : T_1) \leq p(a : T_2)$. These definitions also entail that $p(a : T_1 \wedge T_2) \leq p(a : T_1)$, and $p(a : T_1) \leq p(a : T_1 \vee T_2)$.

We generalize probabilistic meet and join types to probabilities for unbounded conjunctive and disjunctive type judgements, again using the classical equations.

Let $\bigwedge_p(a_0 : T_0, \ldots, a_n : T_n)$ be *the conjunctive probability of judgements* $a_0 : T_0, \ldots, a_n : T_n$. Then $\bigwedge_p(a_0 : T_0, \ldots, a_n : T_n) = \bigwedge_p(a_0 : T_0, \ldots, a_{n-1} : T_{n-1})p(a_n : T_n \mid a_0 : T_0, \ldots, a_{n-1} : T_{n-1})$. If $n = 0$, $\bigwedge_p(a_0 : T_0, \ldots, a_n : T_n) = 1$.

We interpret universal quantification as an unbounded conjunctive probability, which is true if it is vacuously satisfied ($n = 0$) (Paris, 2010).

Let $\bigvee^p(a_0 : T_0, a_1 : T_1, \ldots)$ be the disjunctive probability of judgements $a_0 : T_0, a_1 : T_1, \ldots$. It is computed by $\bigvee^p(a_0 : T_0, \ldots, a_n : T_n) = \bigvee^p(a_0 : T_0, \ldots, a_{n-1} : T_{n-1}) + p(a_n : T_n) - \bigwedge_p(a_0 : T_0, \ldots, a_{n-1} : T_{n-1})p(a_n : T_n \mid a_0 : T_0, \ldots, a_{n-1} : T_{n-1})$. If $n = 0$, $\bigvee^p(a_0 : T_0, \ldots, a_n : T_n) = 0$.

We take existential quantification to be an unbounded disjunctive probability, which is false if it lacks a single non-nil probability instance ($n = 0$).

Conditional Conjunctive Probabilities are computed by $\bigwedge_p(a_0 : T_0, \ldots, a_n : T_n \mid a : T) = \bigwedge_p(a_0 : T_0, \ldots, a_{n-1} : T_{n-1} \mid a : T)p(a_n : T_n \mid a_0 : T_0, \ldots, a_{n-1} : T_{n-1}, a : T))$. If $n = 0$, $\bigwedge_p(a_0 : T_0, \ldots, a_n : T_n \mid a : T) = 1$.

Function Types give, for any types T_1 and T_2, the type $(T_1 \to T_2)$. This is the type of total functions with domain the set of all objects of type T_1 and range included in objects of type T_2. The probability that a function f is of type $(T_1 \to T_2)$ is the probability that everything in its domain is of type T_1 and that everything in its range is of type T_2, and furthermore that everything not in its domain which has some probability of being of type T_1 is *not* in fact of type T_1. $p(f : (T_1 \to T_2)) = \bigwedge_{a \in \mathrm{dom}(f)}^p(a : T_1, f(a) : T_2)(1 - \bigvee_{a \notin \mathrm{dom}(f)}^p(a : T_1))$

Suppose that T_1 is the type of event where there is a flash of lightning and T_2 is the type of event where there is a clap of thunder. Suppose that f maps lightning events to thunder events, and that it has as its domain all events which have been judged to have probability greater than 0 of being lightning events. Let us consider that all the putative lightning events were clear examples of lightning (i.e. judged with probability 1 to be of type T_1) and are furthermore associated by f with clear events of thunder (i.e. judged with probability 1 to be of type T_2). Suppose there were four such pairs of events. Then the probability of f being of type $(T_1 \to T_2)$ is $(1 \times 1)^4$, that is, 1.

Suppose, alternatively, that for one of the four events f associates the lightning event with a silent event, that is, one whose probability of being of T_2 is 0. Then the probability of f being of type $(T_1 \to T_2)$ is $(1 \times 1)^3 \times (1 \times 0) = 0$. One clear counterexample is sufficient to show that the function is definitely not of the type.

In cases where the probabilities of the antecedent and the consequent type judgements are higher than 0, the probability of the entire judgement on the existence of a functional type f will decline in proportion to the size of $\mathrm{dom}(f)$. Assume, for example that there are k elements $a \in \mathrm{dom}(f)$, where for each such a $p(a : T_1) = p(f(a) : T_2) \geq .5$. Every a_i that is added to $\mathrm{dom}(f)$ will reduce the value of $p(f : (T_1 \to T_2))$, even if it yields higher values for $p(a : T_1)$ and $p(f(a) : T_2)$. This is due to the fact that we are treating the probability of $p(f : (T_1 \to T_2))$ as the likelihood of there being a function that is satisfied by all objects in its domain. The larger the domain, the less probable that all elements in it fulfill the functional relation.

We are, then, interpreting a functional type judgement of this kind as a universally quantified assertion over the pairing of objects in $\mathrm{dom}(f)$ and $\mathrm{range}(f)$. The probability of such an assertion is given by the conjunction of assertions corresponding to the co-occurrence of each element a in f's domain as an instance of T_1 with $f(a)$ as an instance of T_2. This probability is the product of the probabilities of these individual assertions.

This seems reasonable, but it only deals with functions whose domain is all objects which have been judged to have some probability, however low, of being of type T_1. Intuitively, functions which leave out some of the objects with lower likelihood of being of type T_1 should also have a probability of being of type $(T_1 \to T_2)$. This factor in the probability is represented by the second element of the product in the formula.

Negation $\neg T$, of type T, is the function type $(T \rightarrow \bot)$, where $\bot$ is a necessarily empty type and $p(\bot) = 0$. It follows from our rules for function types that $p(f : \neg T) = 1$ if $\mathrm{dom}(f) = \emptyset$, that is T is empty, and 0 otherwise.

We also assign probabilities to judgements concerning the (non-)emptiness of a type, $p(T)$. we pass over the details of how we compute the probabilities of such judgements, but we note that our account of negation entails that $p(T \vee \neg T) = 1$, and (ii) $p(\neg\neg T) = p(T)$. Therefore, we sustain classical Boolean negation and disjunction, in contrast to Martin-Löf's (1984) intuitionistic type theory.

Dependent Types are functions from objects to types. Given appropriate arguments as functions they will return a type. Therefore, the account of probabilities associated with functions above applies to dependent types.

Record Types A record in a type system associated with a set of labels is a set of ordered pairs (*fields*) whose first member is a label and whose second member is an object of some type (possibly a record). Records are required to be functional on labels (each label in a record can only occur once in the record's left projection).

A dependent record type is a set of fields (ordered pairs) consisting of a label ℓ followed by T as above. The set of record types is defined by:

1. [], that is the empty set or *Rec*, is a record type. $r : Rec$ just in case r is a record.

2. If T_1 is a record type, ℓ is a label not occurring in T_1, and T_2 is a type, then $T_1 \cup \{\langle \ell, T_2 \rangle\}$ is a record type. $r : T_1 \cup \{\langle \ell, T_2 \rangle\}$ just in case $r : T_1$, $r.\ell$ is defined (ℓ occurs as a label in r) and $r.\ell : T_2$.

3. If T is a record type, ℓ is a label not occuring in T, $\mathcal{T}$ is a dependent type requiring n arguments, and $\langle \pi_1, \ldots, \pi_n \rangle$ is an n-place sequence of paths in T,[4] then $T \cup \{\langle \ell, \langle \mathcal{T}, \langle \pi_1, \ldots, \pi_n \rangle \rangle \rangle\}$ is a record type. $r : T \cup \{\langle \ell, \langle \mathcal{T}, \langle \pi_1, \ldots, \pi_n \rangle \rangle \rangle\}$ just in case $r : T$, $r.\ell$ is defined and $r.\ell : \mathcal{T}(r.\pi_1, \ldots, r.\pi_n)$.

The probability that an object r is of a record type T is given by the following clauses:

1. $p(r : Rec) = 1$ if r is a record, 0 otherwise

2. $p(r : T_1 \cup \{\langle \ell, T_2 \rangle\}) = \bigwedge_p (r : T_1, r.\ell : T_2)$

3. If $\mathcal{T} : (T_1 \rightarrow (\ldots \rightarrow (T_n \rightarrow Type)\ldots))$, then $p(r : T \cup \{\langle \ell, \langle \mathcal{T}, \langle \pi_1, \ldots, \pi_n \rangle \rangle \rangle\}) = \bigwedge_p (r : T, r.\ell : \mathcal{T}(r.\pi_1, \ldots, r.\pi_n) \mid r.\pi_1 : T_1, \ldots, r.\pi_n : T_n)$

[4] In the full version of TTR we also allow absolute paths which point to particular records, but we will not include them here.

3 Compositional Semantics

Montague (1974) determines the denotation of a complex expression by applying a function to an intensional argument (as in $[\![\text{ NP }]\!]([\![\text{ }^\wedge\text{VP }]\!])$). We employ a variant of this general strategy by applying a probabilistic evaluation function $[\![\cdot]\!]_p$ to a categorical (non-probabilistic) semantic value. For semantic categories that are interpreted as functions, $[\![\cdot]\!]_p$ yields functions from categorical values to probabilities. For sentences it produces probability values.

The probabilistic evaluation function $[\![\cdot]\!]_p$ produces a probabilistic interpretation based on a classical compositional semantics. For sentences it will return the probability that the sentence is true. For categories that are interpreted as functions it will return functions from (categorical) interpretations to probabilities. We are not proposing strict compositionality in terms of probabilities. Probabilities are like truth-values (or rather, truth-values are the limit cases of probabilities).

We would not expect to be able to compute the probability associated with a complex constituent on the basis of the probabilities associated with its immediate constituents, any more than we would expect to be able to compute a categorical interpretation entirely in terms of truth-functions and extensions. However, the simultaneous computation of categorical and probabilistic interpretations provides us with a compositional semantic system that is closely related to the simultaneous computation of intensions and extensions in classical Montague semantics.

The following definition of $[\![\cdot]\!]_p$ for a fragment of English is specified on the basis of our probabilistic type system and a non-probabilistic interpretation function $[\![\cdot]\!]$, which we do not give in this version of the paper. (It's definition is given by removing the probability p from the definition below.)

$[\![[_{\text{S}} \text{ S}_1 \text{ and S}_2]]\!]_p = \text{p}(\begin{bmatrix} \text{e}_1{:}[\![\text{ S}_1 \text{ }]\!] \\ \text{e}_2{:}[\![\text{ S}_2 \text{ }]\!] \end{bmatrix})$

$[\![[_{\text{S}} \text{ S}_1 \text{ or S}_2]]\!]_p = \text{p}([\text{e}{:}[\![\text{ S}_1 \text{ }]\!] \vee [\![\text{ S}_2 \text{ }]\!]])$

$[\![[_{\text{S}} \text{ Neg S}]]\!]_p = [\![\text{ Neg }]\!]_p([\![\text{ S }]\!])$

$[\![[_{\text{S}} \text{ NP VP}]]\!]_p = [\![\text{ NP }]\!]_p([\![\text{ VP }]\!])$

$[\![[_{\text{NP}} \text{ Det N}]]\!]_p = [\![\text{ Det }]\!]_p([\![\text{ N }]\!])$

$[\![[_{\text{NP}} \text{ N}_{prop}]]\!]_p = [\![\text{ N}_{prop}]\!]_p$

$[\![[_{\text{VP}} \text{ V}_t \text{ NP}]]\!]_p = [\![\text{ V}_t]\!]_p([\![\text{ NP }]\!])$

$[\![[_{\text{VP}} \text{ V}_i]]\!]_p = [\![\text{ V}_i]\!]_p$

$[\![[_{\text{Neg}} \text{ “it's not true that”}]]\!]_p = \lambda T{:}\mathit{RecType}(\text{p}([\text{e}{:}\neg T]))$

$[\![[_{\text{Det}} \text{ “some”}]]\!]_p = \lambda Q{:}\mathit{Ppty}(\lambda P{:}\mathit{Ppty}(\text{p}([\text{e}{:}\text{some}(Q, P)])))$

$[\![[_{\text{Det}} \text{ “every”}]]\!]_p = \lambda Q{:}\mathit{Ppty}(\lambda P{:}\mathit{Ppty}(\text{p}([\text{e}{:}\text{every}(Q, P)])))$

$[\![[_{\text{Det}} \text{ “most”}]]\!]_p = \lambda Q{:}\mathit{Ppty}(\lambda P{:}\mathit{Ppty}(\text{p}([\text{e}{:}\text{most}(Q, P)])))$

$[\![\ [_{\text{N}}$ "boy"$]\ ]\!]_p = \lambda r{:}[\text{x}{:}\mathit{Ind}](\text{p}([\text{e:boy}(r.\text{x})]))$

$[\![\ [_{\text{N}}$ "girl"$]\ ]\!]_p = \lambda r{:}[\text{x}{:}\mathit{Ind}](\text{p}([\text{e:girl}(r.\text{x})]))$

$[\![\ [_{\text{Adj}}$ "green"$]\ ]\!]_p =$
$\lambda P{:}\mathit{Ppty}(\lambda r{:}[\text{x}{:}\mathit{Ind}](\text{p}([\text{e:green}(r.\text{x},P)])))$

$[\![\ [_{\text{Adj}}$ "imaginary"$]\ ]\!]_p =$
$\lambda P{:}\mathit{Ppty}(\lambda r{:}[\text{x}{:}\mathit{Ind}](\text{p}([\text{e:imaginary}(r.\text{x},P)])))$[5]

$[\![\ [_{\text{N}_{prop}}$ "Kim"$]\ ]\!]_p = \lambda P{:}\mathit{Ppty}(\text{p}(P([\text{x=kim}])))$

$[\![\ [_{\text{N}_{prop}}$ "Sandy"$]\ ]\!]_p = \lambda P{:}\mathit{Ppty}(\text{p}(P([\text{x=sandy}])))$

$[\![\ [_{\text{V}_t}$ "knows"$]\ ]\!]_p =$
$\lambda \mathcal{P}{:}\mathit{Quant}(\lambda r_1{:}[\text{x}{:}\mathit{Ind}](\text{p}(\mathcal{P}(\lambda r_2{:}([\text{e:know}(r_1.\text{x},r_2.\text{x})])))))$

$[\![\ [_{\text{V}_t}$ "sees"$]\ ]\!]_p =$
$\lambda \mathcal{P}{:}\mathit{Quant}(\lambda r_1{:}[\text{x}{:}\mathit{Ind}](\text{p}(\mathcal{P}(\lambda r_2{:}([\text{e:see}(r_1.\text{x},r_2.\text{x})])))))$

$[\![\ [_{\text{V}_i}$ "smiles"$]\ ]\!]_p = \lambda r{:}[\text{x}{:}\mathit{Ind}](\text{p}([\text{e:smile}(r.\text{x})]))$

$[\![\ [_{\text{V}_i}$ "laughs"$]\ ]\!]_p = \lambda r{:}[\text{x}{:}\mathit{Ind}](\text{p}([\text{e:laugh}(r.\text{x})]))$

A probability distribution d for this fragment, based on a set of situations $\mathcal{S}$, is such that:

$p_d(a : \mathit{Ind}) = 1$ if a is kim or sandy[6]

$p_d(s : T) \in [0, 1]$ if $s \in \mathcal{S}$ and T is a ptype

$p_d(s : T) = 0$ if $s \notin \mathcal{S}$ and T is a ptype[7]

$p_d(a : [^\tau P]) = p_d(P([\text{x=}a]))$

$p_d(\text{some}(P, Q)) = p_d([^\tau P] \wedge [^\tau Q])$

$p_d(\text{every}(P, Q)) = p_d([^\tau P] \rightarrow [^\tau Q])$

$p_d(\text{most}(P, Q)) = \min(1, \frac{p_d([^\tau P] \wedge [^\tau Q]}{\theta_{most}\ p_d([^\tau P])})$

The probability that an event e is of the type in which the relation $some$ holds of the properties P and Q is the probability that e is of the conjunctive type $P \wedge Q$. The probability that e is of the $every$ type for P and Q is the likelihood that it instantiates the functional type $P \rightarrow Q$. As we have defined the probabilities associated with functional types in terms of universal quantification (an unbounded conjunction of the pairings between the elements of the domain P of the function and its range Q), this definition sustains the desired reading of *every*. The likelihood that e is of the type $most$ for P and Q is the likelihood that e is of type $P \wedge Q$, factored by the product of the contextually determined parameter θ_{most} and the likelihood that e is of type P, where this fraction is less than 1, and 1 otherwise.

Consider a simple example.

$[\![\ [_{\text{S}}\ [_{\text{NP}}\ [_{\text{N}_{prop}}$ Kim]] $[_{\text{VP}}\ [_{\text{V}_i}$ smiles]]] $]\!]_p =$
$\lambda P{:}\mathit{Ppty}(\text{p}(P([\text{x=kim}])))(\lambda r{:}[\text{x}{:}\mathit{Ind}]([\text{e:smile}(r.\text{x})])) =$
$p(\lambda r{:}[\text{x}{:}\mathit{Ind}]([\text{e:smile}(r.\text{x})])([\text{x=kim}])) =$
$p([\text{e:smile(kim)}])$

Suppose that $p_d(s_1{:}\text{smile(kim)}) = .7$, $p_d(s_2{:}\text{smile(kim)}) = .3$, $p_d(s_3{:}\text{smile(kim)}) = .4$, and there are no other situations s_i such that $p_d(s_i{:}\text{smile(kim)}) > 0$. Furthermore, let us assume that these probabilities are independent of each other, that is, $p_d(s_3{:}\text{smile(kim)}) = p_d(s_3{:}\text{smile(kim)} \mid s_1{:}\text{smile(kim)}, s_2{:}\text{smile(kim)})$ and so on. Then

$p_d(\text{smile(kim)}) =$
$\bigvee^p_d(s_1 : \text{smile(kim)}, s_2 : \text{smile(kim)}, s_3 : \text{smile(kim)}) =$
$\bigvee^p_d(s_1 : \text{smile(kim)}, s_2 : \text{smile(kim)}) + .4 - .4 \bigvee^p_d(s_1 : \text{smile(kim)}, s_2 : \text{smile(kim)}) =$
$(.7 + .3 - .7 \times .3) + .4 - .4(.7 + .3 - .7 \times .3) =$
$.874$

This means that $p_d([\text{e:smile(kim)}]) = .874$. Hence $[\![\ [_{\text{S}}\ [_{\text{NP}}\ [_{\text{N}_{prop}}$ Kim]] $[_{\text{VP}}\ [_{\text{V}_i}$ smiles]]] $]\!]_{p_d} = .874$ (where $[\![\ \alpha\ ]\!]_{p_d}$ is the result of computing $[\![\ \alpha\ ]\!]_p$ with respect to the probability distribution d).

Just as for categorical semantics, we can construct type theoretic objects corresponding to probabilistic judgements. We call these *probabilistic Austinian propositions*. These are records of type

$$\begin{bmatrix} \text{sit} & : & \mathit{Sit} \\ \text{sit-type} & : & \mathit{Type} \\ \text{prob} & : & [0,1] \end{bmatrix}$$

where [0,1] is used to represent the type of real numbers between 0 and 1. They assert that the probability that a situation s is of type $Type$ is the value of $prob$.

The definition of $[\![\ \cdot\ ]\!]_p$ specifies a compositional procedure for generating an Austinian proposition (record) of this type from the meanings of the syntactic constituents of a sentence.

4 An Outline of Semantic Learning

We outline a schematic theory of semantic learning on which agents acquire classifiers that form the basis for our probabilistic type system. For simplicity and ease of presentation we take these to be Naive Bayes classifiers, which an agent acquires from observation. In future developments of this theory we will seek to extend the approach to Bayesian networks (Pearl, 1990).

We assume that agents keep records of observed situations and their types, modelled as probabilistic Austinian propositions. For example, an observation of a man running might yield the following Austinian proposition for some $a{:}\mathit{Ind}$, $s_1{:}\text{man}(a)$, $s_2{:}\text{run}(a)$:

[5] Notice that we characterize adjectival modifiers as relations between records of individuals and properties. We can then invoke subtyping to capture the distinction between intersective and non-intersective modifier relations.

[6] This seems an intuitive assumption, though not a necessary one.

[7] Again this seems an intuitive, though not a necessary assumption.

$$\left[\begin{array}{lcl} \text{sit} & = & \left[\begin{array}{lcl} \text{ref} & = & a \\ \text{c}_{\text{man}} & = & s_1 \\ \text{c}_{\text{run}} & = & s_2 \end{array}\right] \\ \text{sit-type} & = & \left[\begin{array}{lcl} \text{ref} & : & \textit{Ind} \\ \text{c}_{\text{man}} & : & \text{man(ref)} \\ \text{c}_{\text{run}} & : & \text{run(ref)} \end{array}\right] \\ \text{prob} & = & 0.7 \end{array}\right]$$

An agent, A, makes judgements based on a finite string of probabilistic Austinian propositions, $\mathfrak{J}$, corresponding to prior judgements held in memory. For a type, T, $\mathfrak{J}_T$ represents that set of Austinian propositions j such that j.sit-type $\sqsubseteq T$. If T is a type and $\mathfrak{J}$ a finite string of probabilistic Austinian propositions, then $||\ T\ ||_{\mathfrak{J}}$ represents the sum of all probabilities associated with T in $\mathfrak{J}$ ($\sum_{j \in \mathfrak{J}_T} j.prob$). $\mathcal{P}(\mathfrak{J})$ is the sum of all probabilities in $\mathfrak{J}$ ($\sum_{j \in \mathfrak{J}} j.prob$).

We use $\text{prior}_{\mathfrak{J}}(T)$ to represent the prior probability that anything is of type T given $\mathfrak{J}$, that is $\frac{||T||_{\mathfrak{J}}}{\mathcal{P}(\mathfrak{J})}$ if $\mathcal{P}(\mathfrak{J}) > 0$, and 0 otherwise.

$p_{A,\mathfrak{J}}(s : T)$ denotes the probability that agent A assigns with respect to prior judgements $\mathfrak{J}$ to s being of type T. Similarly, $p_{A,\mathfrak{J}}(s : T_1 \mid s : T_2)$ is the probability that agent A assigns with respect to prior judgements $\mathfrak{J}$ to s being of type T_1, given that A judges s to be of type T_2.

When an agent A encounters a new situation s and considers whether it is of type T, he/she uses probabilistic reasoning to determine the value of $p_{A,\mathfrak{J}}(s : T)$. A uses conditional probabilities to calculate this value, where A computes these conditional probabilities with the equation $p_{A,\mathfrak{J}}(s : T_1 \mid s : T_2) = \frac{||T_1 \wedge T_2||_{\mathfrak{J}}}{||T_2||_{\mathfrak{J}}}$, if $||\ T_2\ ||_{\mathfrak{J}} \neq 0$. Otherwise, $p_{A,\mathfrak{J}}(s : T_1 \mid s : T_2) = 0$.

This is our type theoretic variant of the standard Bayesian formula for conditional probabilities: $p(A \mid B) = \frac{|A \& B|}{|B|}$. But instead of counting categorical instances, we sum the probabilities of judgements. This is because our "training data" is not limited to categorical observations. Instead it consists of probabilistic observational judgements that situations are of particular types.[8]

Assume that we have the following types:

$$\text{T}_{\text{man}} = \left[\begin{array}{lcl} \text{ref} & : & \textit{Ind} \\ \text{c}_{\text{man}} & : & \text{man(ref)} \end{array}\right] \text{ and}$$

$$\text{T}_{\text{run}} = \left[\begin{array}{lcl} \text{ref} & : & \textit{Ind} \\ \text{c}_{\text{run}} & : & \text{run(ref)} \end{array}\right]$$

[8] As a reviewer observes, by using an observer's previous judgements for the probability of an event being of a particular type, as the prior for the rule that computes the probability of a new event being of that type, we have, in effect, compressed information that properly belongs in a Bayesian network into our specification of a naive Bayesian classifier. This is a simplification that we adopt here for ease of exposition. In future work, we will characterise classifier learning through full Bayesian networks.

Assume also that $\mathfrak{J}_{T_{\text{man}} \wedge T_{\text{run}}}$ has three members, corresponding to judgements by A that a man was running in three observed situations s_1, s_3, and s_4, and that these Austinian propositions have the probabilities 0.6, 0.6. and 0.5 respectively.

Take $\mathfrak{J}_{T_{\text{man}}}$ to have five members corresponding to judgements by A that there was a man in $s_1, \ldots, s_5$, and that the Austinian propositions assigning T_{man} to $s_1, \ldots, s_5$ all have probability 0.7. Given these assumptions, the conditional probability that A will assign on the basis of $\mathfrak{J}$ to someone runs, given that he is a man is $p_{A,\mathfrak{J}}(r : \text{T}_{\text{run}} \mid r : \text{T}_{\text{man}}) = \frac{||T_{\text{man}} \wedge T_{\text{run}}||_{\mathfrak{J}}}{||T_{\text{man}}||_{\mathfrak{J}}} = \frac{0.6+0.6+0.5}{0.7+0.7+0.7+0.7+0.7} = .486$

We use conditional probabilities to construct a Naive Bayes classifier. A classifies a new situation s based on the prior judgements $\mathfrak{J}$, and whatever evidence A can acquire about s. This evidence has the form $p_{A,\mathfrak{J}}(s : T_{e_1})$, ..., $p_{A,\mathfrak{J}}(s : T_{e_n})$, where $T_{e_1}, \ldots, T_{e_n}$ are the *evidence types*. The Naive Bayes classifier assumes that the evidence is independent, in that the probability of each piece of evidence is independent of every other piece of evidence.

We first formulate Bayes' rule of conditional probability. This rule defines the conditional probability of a conclusion $r : T_c$, given evidence $r : T_{e_1}, r : T_{e_2}, \ldots, r : T_{e_n}$, in terms of conditional probabilities of the form $p(s_i : T_{e_i} \mid s_i : T_c)$, $1 \leq i \leq n$, and *priors* for conclusion and evidence:

$$p_{A,\mathfrak{J}}(r : T_c \mid r : T_{e_1}, \ldots, r : T_{e_n}) = \text{prior}_{\mathfrak{J}}(T_c) \frac{\frac{||T_{e_1} \wedge T_c||_{\mathfrak{J}}}{||T_c||_{\mathfrak{J}}} \ldots \frac{||T_{e_n} \wedge T_c||_{\mathfrak{J}}}{||T_c||_{\mathfrak{J}}}}{\text{prior}_{\mathfrak{J}}(T_{e_1}) \ldots \text{prior}_{\mathfrak{J}}(T_{e_n})}$$

The conditional probabilities are computed from observations as indicated above. The rule of conditional probability allows the combination of several pieces of evidence, without requiring previous observation of a situation involving all the evidence types.

We formulate a Naive Bayes classifier as a function from evidence types $T_{e_1}, T_{e_2}, \ldots, T_{e_n}$ (i.e. from a record of type $T_{e_1} \wedge T_{e_2} \wedge \ldots \wedge T_{e_n}$) to conclusion types $T_{c_1}, T_{c_2}, \ldots, T_{c_m}$. The conclusion is a disjunction of one or more $T \in \{T_{c_1}, T_{c_2}, \ldots, T_{c_m}\}$, where m ranges over all possible non-disjunctive conclusions distinguished by the classifier. This function is specified as follows.

$$\kappa : (T_{e_1} \wedge \ldots \wedge T_{e_n}) \rightarrow (T_{c_1} \vee \ldots \vee T_{c_m}) \text{ such that } \kappa(r) = (\bigvee \underset{T \in \langle T_{c_1}, \ldots, T_{c_m} \rangle}{\text{argmax}} p_{A,\mathfrak{J}}(r : T \mid r : T_{e_1}, \ldots, r : T_{e_n})$$

The classifier returns the type T which maximises the conditional probability of $r : T$ given

the evidence provided by r. The argmax operator here takes a sequence of arguments and a function and yields a sequence containing the arguments which maximise the function (if there are more than one).

The classifier will output a disjunction in case both possibilities have the same probability. The $\bigvee$ operator takes a sequence and returns the disjunction of all elements of the sequence.

In addition to computing the conclusion which receives the highest probability given the evidence, we also want the *posterior* probability of the judgement above, i.e. the probability of the judgement in light of the evidence. We obtain the non-normalised probabilities ($p^{\text{nn}}_{A,\mathfrak{J}}$) of the different possible conclusions by factoring in the probabilities of the evidence:

$p^{\text{nn}}_{A,\mathfrak{J}}(r : \kappa(r)) =$
$\sum_{T \in \bigvee^{-1} \kappa(r)} p_{A,\mathfrak{J}}(r : T \mid r : T_{e_1}, \ldots, r : T_{e_n}) p_{A,\mathfrak{J}}(r : T_{e_1}) \ldots p_{A,\mathfrak{J}}(r : T_{e_n})$
where $\bigvee^{-1}$ is the inverse of $\bigvee$, i.e. a function that takes a disjunction and returns the set of disjuncts.

We then take the probability of $r : \kappa(r)$ and normalise over the sum of the probabilities of all the possible conclusions. This gives us the normalised probability of the judgement resulting from classification $p(r : \kappa(r)) = \frac{\text{p}^{\text{nn}}_{A,\mathfrak{J}}(r:\kappa(r))}{\sum_{1 \leq i \leq m} \text{p}^{\text{nn}}_{A,\mathfrak{J}}(r:T_{c_i})}$.

However, since the probabilities of the evidence are identical for all possible conclusions, we can ignore them and instead compute the normalised probability with the following equation (where m ranges over all possible non-disjunctive conclusions distinguished by the classifier, as above).

$$p_{A,\mathfrak{J}}(r : \kappa(r)) = \frac{\sum_{T \in \bigvee^{-1} \kappa(r)} p_{A,\mathfrak{J}}(r:T \mid r:T_{e_1}, \ldots, r:T_{e_n})}{\sum_{1 \leq i \leq m} p_{A,\mathfrak{J}}(r:T_{c_i} \mid r:T_{e_1}, \ldots, r:T_{e_n})}$$

The result of classification can be represented as an Austinian proposition

$$\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & \kappa(s) \\ \text{prob} & = & p_{A,\mathfrak{J}}(s : \kappa(s)) \end{bmatrix}$$

which A adds to $\mathfrak{J}$ as a result of observing and classifying s, and is thus made available for subsequent probabilistic reasoning.

5 Conclusions and Future Work

We have presented a probabilistic version of a rich type theory with records, relying heavily on classical equations for types formed with meet, join, and negation. This has permitted us to sustain classical equivalences and Boolean negation for complex types within an intensional type theory. We have replaced the truth of a type judgement with the probability of it being the case, and we have applied this approach to judgements that a situation if of type T.

Our probabilistic formulation of a rich type theory with records provides the basis for a compositional semantics in which functions apply to categorical semantic objects in order to return either functions from categorical interpretations to probabilistic judgements, or, for sentences, to probabilistic Austinian propositions. One of the interesting ways in which this framework differs from classical model theoretic semantics is that the basic types and type judgements at the foundation of the type system correspond to perceptual judgements concerning objects and events in the world, rather than to entities in a model and set theoretic constructions defined on them.

We have offered a schematic view of semantic learning. On this account observations of situations in the world support the acquisition of naive Bayesian classifiers from which the basic probabilistic types of our type theoretical semantics are extracted. Our type theory is, then, the interface between observation-based learning of classifiers for objects and the situations in which they figure on one hand, and the computation of complex semantic values for the expressions of a natural language from these simple probabilistic types and type judgements on the other. Therefore our general model of interpretation achieves a highly integrated bottom-up treatment of linguistic meaning and perceptually-based cognition that situates meaning in learning how to make observational judgements concerning the likelihood of situations obtaining in the world.

The types of our semantic theory are intensional. They constitute ways of classifying situations, and they cannot be reduced to set of situations. The theory achieves fine-grained intensionality through a rich and articulated type system, where the foundation of this system is anchored in perceptual observation.

The meanings of expressions are acquired on the basis of speakers' experience in the application of classifiers to objects and events that they encounter. Meanings are dynamic and updated in light of subsequent experience.

Probability is distributed over alternative situation types. Possible worlds, construed as maximal consistent sets of propositions (ultrafilters in a proof theoretic lattice of propositions) play no role in this framework.

Bayesian reasoning from observation provides the incremental basis for learning and refining predicative types. These types feed the combinatorial semantic procedures for interpreting the sentences of a natural language.

In future work we will explore implementations of our learning theory in order to study the viability of our probabilistic type theory as an interface between perceptual judgement and compositional semantics. We hope to show that, in addition to its cognitive and theoretical interest, our proposed framework will yield results in robotic language learning, and dialogue modelling.

Acknowledgments

We are grateful to two anonymous reviewers for very helpful comments on an earlier draft of this paper. We also thank Alex Clark, Jekaterina Denissova, Raquel Fernández, Jonathan Ginzburg, Noah Goodman, Dan Lassiter, Michiel van Lambalgen, Poppy Mankowitz, Aarne Ranta, and Peter Sutton for useful discussion of ideas presented in this paper. Shalom Lappin's participation in the research reported here was funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK, and a grant from the Wenner-Gren Foundations. We also gratefully acknowledge the support of Vetenskapsrådet, project 2009-1569, Semantic analysis of interaction and coordination in dialogue (SAICD); the Department of Philosophy, Linguistics, and Theory of Science; and the Centre for Language Technology at the University of Gothenburg.

References

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. Bradford Books. MIT Press, Cambridge, Mass.

I. Beltagy, C. Chau, G. Boleda, D. Garrette, K. Erk, and R. Mooney. 2013. Montague meets markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics, Vol. 1*, pages 11–21. Association of Computational Linguistics, Atlanta, GA.

R. Carnap. 1947. *Meaning and Necessity*. University of Chicago Press, Chicago.

A. Clark and S. Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Chichester, West Sussex, and Malden, MA.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier BV, 271–323. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

C. Fox and S. Lappin. 2010. Expressiveness and complexity in underspecified semantics. *Linguistic Analysis, Festschrift for Joachim Lambek*, 36:385–417.

J. Halpern. 2003. *Reasoning About Uncertainty*. MIT Press, Cambridge MA.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

A.N. Kolmogorov. 1950. *Foundations of Probability*. Chelsea Publishing, New York.

Per Martin-Löf. 1984. *Intuitionistic Type Theory*. Bibliopolis, Naples.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven. ed. and with an introduction by Richmond H. Thomason.

C. Papadimitriou. 1995. *Computational Complexity*. Addison-Wesley Publishing Co., Readin, MA.

J. Paris. 2010. Pure inductive logic. Winter School in Logic, Guangzhou, China.

J. Pearl. 1990. Bayesian decision methods. In G. Shafer and J. Pearl, editors, *Readings in Uncertain Reasoning*, pages 345–352. Morgan Kaufmann.