Bayesian Inference Semantics: A Modelling System and A Test Suite

Jean-Philippe Bernardy **Rasmus Blanck** **Stergios Chatzikyriakidis**
Shalom Lappin **Aleksandre Maskharashvili**
University of Gothenburg
firstname.lastname@gu.se

Abstract

We present BIS, a Bayesian Inference Semantics, for probabilistic reasoning in natural language. The current system is based on the framework of Bernardy et al. (2018), but departs from it in important respects. BIS makes use of Bayesian learning for inferring a hypothesis from premises. This involves estimating the probability of the hypothesis, given the data supplied by the premises of an argument. It uses a syntactic parser to generate typed syntactic structures that serve as input to a model generation system. Sentences are interpreted compositionally to probabilistic programs, and the corresponding truth values are estimated using sampling methods. BIS successfully deals with various probabilistic semantic phenomena, including frequency adverbs, generalised quantifiers, generics, and vague predicates. It performs well on a number of interesting probabilistic reasoning tasks. It also sustains most classically valid inferences (instantiation, de Morgan's laws, etc.). To test BIS we have built an experimental test suite with examples of a range of probabilistic and classical inference patterns.

1 Introduction

On a traditional view of inference, the entailment relation between the premises of an argument and its conclusion holds *iff* the argument is logically valid in a proof or model theory. More recently, computational approaches to entailment in natural text, such as Recognising Textual Entailment (RTE, Dagan et al. (2009)) have attempted to capture inferences that depend on lexical meaning and real world knowledge, as well as logical structure. In the latter sorts of inference, the conclusions often follow from the premises within a certain range of probability values.

In this paper we present Bayesian Inference Semantics (BIS), a probabilistic semantics for natural language that assigns probability values, rather than Boolean truth-values, to sentences. The probability of a sentence is the likelihood that an idealised speaker, as represented by our model, would accept the assertion that it expresses. Our framework builds on the approach proposed by Bernardy et al. (2018). It is Bayesian in that it constructs models in which asserted constraints provide Bayesian evidence that models use to determine whether objects satisfy particular properties.

Objects are represented as vectors in a model space S, and properties are subspaces in S. Satisfaction of a property is expressed as membership in the corresponding subspace of S. The probability density over the space of possible situations corresponds to the *a priori* density of objects in these subspaces, and is specified through Bayesian priors. The system leverages the probabilistic functional programming language WebPPL (Goodman and Stuhlmüller, 2014) to evaluate Bayesian posteriors. English sentences are parsed using Ranta's Grammatical Framework (GF, http://www.grammaticalframework.org/, 2004), and the parses are compositionally mapped into interpretations within BIS's probabilistic models.

We apply BIS to inferences, most of which are probabilistic in nature, and so closely related to RTE concerns. We have constructed a test suite of 78 inferences on which we have developed and tested BIS. The premises in each argument provide Bayesian evidence for the models in which the inference is interpreted, and its conclusion is assigned a (posterior) probability value.[1]

In Section 2 we describe BIS. We explain the syntax-model interface that our GF parses provide, and we characterise how our models are constructed. The models employ Monte Carlo

[1] Our test set, and the code for BIS are available at https://github.com/GU-CLASP/bbclm2019.

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 263–272

Markov Chain (MCMC) sampling[2] on objects to estimate membership in the property classes that correspond to the predicates identified in GF parses of our input sentences.

Section 3 presents our inference system and our test set. BIS currently handles a range of generalised quantifiers, sentential and VP negation, modal and temporal adverbs, measure and comparative adjectives, common nouns, and a variety of logical connectives. It treats VPs and common nouns as monadic predicates. We show how BIS handles both probabilistic and logically valid inferences over a series of examples from the test set. We specify the coverage that the system currently achieves for this set. We have designed BIS to capture inferences involving gradable predicates like *tall*, where the application of the predicate is clear for upper and lower bound cases, but increasingly indeterminate for intermediate instances between these points. BIS handles arguments in which neither a predicate nor its contrary apply. It also covers both wide and narrow scope readings of certain quantifiers.

We discuss other approaches to probabilistic semantic inference in Section 4. Finally, in Section 5 we identify the issues that we plan to take up in future work, and state our conclusions.

2 System Description

2.1 Overview

Our system for probabilistic semantics is comprised of three phases: (i) parsing, using the GF tool, (ii) compositional Montegovian Semantics, written in Haskell, and (iii) computation of entailment probability.

Our syntax is encoded in the Grammatical Framework (GF) formalism. GF converts a syntactically well-formed sentence into an abstract syntax tree, which is mapped to semantics. The adequacy of the mapping is guaranteed by using the same types in the GF abstract syntax as in the Haskell semantics. The main constructions are described below.

Our semantics blends aspects of Montague semantics, vector space models, and Bayesian inference. It adopts the main ideas of Bernardy et al. (2018), which we summarise here.

A sentence is interpreted as a probabilistic program returning a Boolean value. Individuals are represented as (probabilistic) vectors. Other syntactic categories are mapped to functional types, following the Montague Grammar paradigm. BIS evaluates the validity of an inference as follows. It expresses the priors as a distribution over individuals and predicates. The premises of the inference impose conditions on the model that correspond to Bayesian observations. We then compute the truth-value of the conclusion using posterior distributions for input variables, yielding a Bernouilli distribution. Its expected value (a real number between 0 and 1) corresponds to a probabilistic measure of entailment. Fig. 1 gives a schematic view of BIS's architecture.

This value can be computed symbolically, for example by using the precise semantics for probabilistic programming of Borgström et al. (2013). However symbolic expressions may contain intractable integrals. Therefore we resort to approximating the result by using MCMC sampling, as described by Goodman et al. (2008), and implemented in their WebPPL tool.

2.2 Basics

Consider the sentence "John is a musician". Our GF grammar parses this sentence as:
$CltoS\ Pos\ (S1\ John\ (Bare\ (IsA\ Musician)))$.
We briefly review the combinators used above. $CltoS$ is of type $Pol \rightarrow Cl \rightarrow S$: it produces a declarative sentence from a polarity (positive or negative) and a declarative clause. $S1$ is of type $NP \rightarrow AVP \rightarrow Cl$, taking a noun phrase and a (possibly modified) verb phrase to return a declarative clause. Here, AVP is understood as VP phrase that might have been modified by a modal adverb. There is no modifier in this example, which is signaled by the combinator $Bare$, of type $VP \rightarrow AVP$. The types of the remaining constants are $IsA : CN \rightarrow VP$, $John : NP$, and $Musician : CN$.

Because the types are the same in Haskell and in GF, any abstract syntax tree given by GF will be a well-typed Haskell expression. To obtain a complete semantics, our model treats $John$ and $Musician$ as random representatives of their respective classes, and they are sampled accordingly. Then the truth value of the sentence is evaluated (there is no premise in this case).

$$
\begin{aligned}
&modelJohnMusician = \mathbf{do}\\
&\ john \leftarrow newInd\\
&\ musician \leftarrow newPred\\
&\ return\ (cltoS\ pos\ (s1\ john\ (bare\ (isA\ musician))))
\end{aligned}
$$

[2]For detailed discussion of MCMC see Brooks (1998); Roberts and Rosenthal (2004).

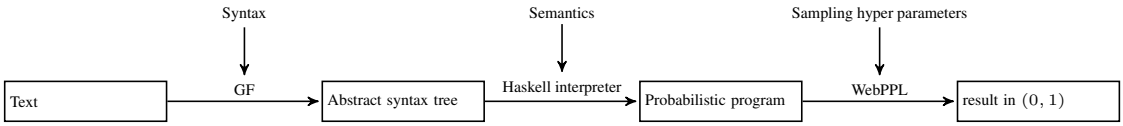

Figure 1: Phases in our system

Running the model, with our implementation, gives the following result:

```
false : 0.67    true : 0.33
```

In the absence of further information, an arbitrary predicate has a chance of (around) 0.33 to hold of an arbitrary individual. This number follows from the way that we model predicates, as described in Section 2.3.

To record assumptions about individuals and predicates, we use the *observe* primitive of Borgström et al. (2013), which ensures that a given proposition holds, and so influences the posteriors. In the MCMC implementation, if the argument to an *observe* call is false, then the corresponding choice of parameters is not retained in the final computation of posteriors. So, for instance, assume that we add the premise that "Most people are musicians" to the earlier example. The premise is parsed as $CltoS\ Pos\ (S1\ (QNP\ Most\ Person)\ (Bare\ (IsA\ Musician)))$, where $QNP{:}Quant \rightarrow CN \rightarrow NP$ and $Most : Quant$. The semantics is:

$$
\begin{array}{l}
model JohnMusicianMost = \mathbf{do} \\
\ john \leftarrow newInd \\
\ musician \leftarrow newPred \\
\ observe\ (cltoS\ pos\ (s1\ (qNP\ most\ person) \\
\qquad\qquad (bare\ (isA\ musician)))) \\
\ return\ (cltoS\ pos\ (s1\ john\ (bare\ (isA\ musician))))
\end{array}
$$

The premise raises the estimated probability value of the conclusion to

```
true : 0.834    false : 0.166
```

2.3 Predicates and their negation

Our basic assumption is that (in the absence of other information) individuals are drawn from a multi-variate normal distribution of dimension k, with a zero mean vector and a unit covariance matrix, where k is a hyperparameter of the system. Any *logical* predicate is represented as a subspace of all individuals. We make the additional simplifying assumption that every atomic lexical predicate p is represented by three components: (1) A vector v_p. The projection of any individual x onto this direction ($x \cdot v_p$) corresponds to the degree to which x exhibits the characteristics corresponding to p. (2) A threshold θ_p^+, such that if $x \cdot v_p > \theta_p^+$ then x is (probabilistically) considered to satisfy p. (3) Another threshold θ_p^-, such that if $x \cdot v_p < \theta_p^-$, then the contrary of the predicate (expressed as VP negation) applies. This procedure allows BIS to express the indeterminacy attached to both measure and non-measure predicates in cases at the border of a classifier, in a fully uniform way.

The vectors v_p are sampled from the same multi-variate normal distribution as individuals, but unlike individuals, these vectors are normalised. Both thresholds are sampled in a standard normal distribution. At all times, we maintain a positive gap between these thresholds: $\theta_p^- < \theta_p^+$. Hence, in our system, the law of the excluded middle does not hold at the linguistic level, since for a given individual x and a predicate p we may well have $\theta_p^- < x \cdot v_p < \theta_p^+$. For example, it is possible that neither "John is a musician" nor "John isn't a musician" apply. Notice that this is not a case of epistemic uncertainty. Rather, in this example, John does not clearly satisfy the property of being a musician and, at the same time, he cannot be regarded as a non-musician. This state of affairs is illustrated in Fig. 2.

2.4 Adjectives and Measure Predicates

Adjectives behave like other predicates. They come with a direction and two thresholds, allowing for the proposition "It is not the case that John is tall, and it is not the case that John isn't tall" to hold in some models.

BIS supports reasoning with units of measures, as in "John is 6 feet tall". We can also express height in various units of measures ("John is 180 cm tall"). To capture the scaling needed for *any* metric we introduce an additional layer of interpretation, which corresponds to units of measure. Each unit of measure u is represented by a pair of a factor f_u and a bias b_u, both drawn from normal distributions. This yields a transformation $t_u(x) = f_u x + b_u$. The numbers provided in the input ("6", "180") are then compared with the transformed measure predicates corresponding to the adjective. (In our example $t_{feet}(john \cdot v_{tall}) = 6$.)

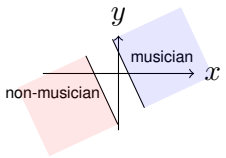

Figure 2: A representation of the predicate "musician" and its negation. The blue and red areas indicate the corresponding subspaces.

This allows BIS to simultaneously infer posterior distributions for individuals, predicates and units of measures.

2.5 Quantifiers

BIS uses the same mechanism for handling quantifiers as Bernardy et al. (2018) do. To interpret a sentence such as "Most musicians are tall", it runs an inner instance of an inference corresponding to "If x is a musician, then x is tall". Then, it imposes, as a posterior of the outer model, the condition that the probabilistic evaluation of the inner inference is higher than a given threshold θ_{most}.

We allow a sentence to contain several generalised quantifiers, such as in "Most bass players are taller than most guitarists", and there are two possibilities to consider when implementing support for this. The first is to nest one application of the above procedure within another. This gives an inner model "If x is a bass player, then x is taller than most guitarists", and an inner-inner inference problem "If y is a guitarist, then x is taller than y". The other is to use a single inner model, with simultaneous quantification over all variables: "If x is a bass players and y is a guitarist, then x is taller than y". The first interpretation is inefficient. Each inner model demands a separate MCMC sampling. When running two-levels of sampling the speed is inversely propositional to the square of samples used. But the second interpretation is not quite correct. The threshold that is being used can only commutatively compound the thresholds of both quantifiers that are used. Therefore, the model would not distinguish between the sentences "Every bass player is taller than most guitarists" and "Most bass players are taller than every guitarist", although their semantics are distinct. For this reason, we opt for the inefficient but precise first procedure, even while we recognise that the second option is a viable way of doing rough grained estimated reasoning.

Finally, we note that it is very inefficient to use the inner-instance sampling method to ensure that a model satisfies sentences containing the universal quantifier, such as "Every musician is a logician". The priors must be set in a particular way to ensure that the subspace of "musicians" is included in that of "logicians". The defining vectors must be exactly parallel. Therefore sampling converges slowly. To deal with this problem we instead impose on the model the requirement that the cosine similarity between the vectors corresponding to "musician" and "logician" is greater than 0.99, and that the threshold $\theta^{+}_{musician}$ is greater than $\theta^{+}_{logician}$. These conditions produce a near-perfect containment of "musician" in "logician".

3 Test Suite

In order to illustrate BIS's coverage we have constructed a test suite. We construct a test suite rather than use any of the existing test suites for inference because all of the latter (e.g. the FraCaS test suite (Cooper et al., 1996), RTE (Dagan et al., 2006), and SNLI (Bowman et al., 2015)) are not designed to assess probabilistic inference, but categorical entailment. They are annotated for three-way (YES, NO, UNK), or binary (YES, NO) values for entailment. In the latter case, the categories NO and UNK of the three-way task are collapsed into a single category. By contrast, we are interested in capturing the full distribution of probability over an inference. Our suite includes 78 examples, each with one or more premises followed by a conclusion. The examples are annotated with respect to the semantic phenomena that figure in the inference. Here is the first example from our test suite:

(T1) P1. Every violinist is a musician.
P2. Musicians generally read music.
H. If John is a violinist, then John reads music.
Label: QUANTIFIER, MODAL ADVERB

Below, we describe several phenomena that an adequate natural language inference system ought to capture. These are particularly important for semantic frameworks designed to handle probabilistic reasoning. While most examples in the current version of the test suite involve probabilistic reasoning, others are classically valid entailments.

We briefly comment on the current state of the art of our system with respect to each of the examples that we present. It is important to note that none of the inferences in our test suite turn on real world knowledge, beyond the information

contained in the premises. This is due to the fact that our models estimate the likelihood of an inference as the conditional probability of the conclusion, given the premises. The premises serve as priors on the models generated to evaluate the conclusion. The models sample the possible relations among the individuals and properties that interpret the NPs and predicates in the conclusion, given only the restrictions imposed by the relations among the individuals and the properties that interpret the statements in the premises.

All the phenomena presented in this section (as seen in examples under the label (Tn), where n is a number of an example in the test suite) are supported by our system.

3.1 Relation to classical logic

Although our main goal is not to embed a particular logic into our system, it is useful to know the relation between our system and a specified logic. This allows us to evaluate whether our system can be used in situations where precise reasoning is required.

BIS nearly supports full classical propositional logic, using the sentential connectives "and", "or", "if ... then ...", together with "it is not the case that". To see that it goes beyond intuitionistic propositional logic we check that it validates the law of the excluded middle, de Morgan's laws, and Peirce's formula. However, BIS does not sustain *reductio ad absurdum* as a rule of inference. The system evaluates an inference by constructing a model for the premises of the inference and evaluating the hypothesis in that model. This means that in arguments where the premises are inconsistent (as would be the case in an attempt to use *reductio*) the system fails to construct such a model. It would not evaluate the hypothesis in that model, yielding no result at all. Therefore the consequence relation is not monotone, as the addition of an extra inconsistent premise makes the computation diverge. By contrast, the system assigns probability 1 to $\models A \wedge \neg A \rightarrow B$ for any choice of A and B.

VP-level negation does not interact with sentential negation in the way that one might expect. GF only provides binary clausal polarity, whereas our implementation of predicates requires a many-valued logic. Therefore, VP negation in examples such as "John isn't a guitarist" rules out both that John is a guitarist, and that John is in the undecided area between $\theta^{-}_{guitarist}$ and $\theta^{+}_{guitarist}$. Hence, "John isn't a guitarist" implies "It is not the case that John is a guitarist", but the converse implication does not hold. As a consequence of our treatment of VP negation, the universal quantifier ("all") and the existential quantifier ("some") are not interdefinable, as they are in classical logic. So, for example, "All musicians read music" implies "It is not the case that some musicians don't read music", but not the other way around.

Instantiation: Instantiation, one of the main inference rules in the Aristotelian syllogism, is supported in our system. The following test suite example is evaluated as *true* with probability 1.

(T31) P1. All intermediate logic students are Stones fans.
P2. John is an intermediate logic student.
H. John is a Stones fan.
Label: INSTANTIATION

Chains of universal affirmatives: The system does not perform very well on examples that chain universal quantifiers together to form valid FOL inferences. Consider the following:

(T76) P1. All violinists are musicians.
P2. All musicians read music.
H. All violinists read music.
Label: QUANTIFIERS, FOL VALIDITY

Here, we would expect a probability of 1 for the conclusion, but the actual result is slightly lower. The reason for this is that our system evaluates the universal quantifier by measuring the cosine similarity between the corresponding vectors (> 0.99) as well as comparing the thresholds θ^{+}_{p} for the two predicates. Even if the cosine similarity between the vectors corresponding to "violinist" and "musician" is close to 1, and the one between "musician" and "read music" is also, this does not imply that the cosine similarity between "violinist" and "read music" is 1.

By contrast, the following example is assigned a probability close to 1, because the percentage determiner is treated as a generalised quantifier, triggering an inner-model inference.

(T77) P1. All violinists are musicians.
P2. All musicians read music.
H. 99 percent of violinists read music.
Label: QUANTIFIERS, PERCENTAGE DETERMINER

Higher-order case: Universally quantified sentences connected via implications Consider an example with more complex cases of conditionals of the form "if X then Y", where X and Y are quantified assertions.

(T78) P1. Every guitarist is a logician.
P2. If every guitarist is a logician, then every musician reads music.
P3. John is a musician.
H. John reads music.
Label: IMPLICATION, QUANTIFIER, PROPER NAME

BIS performs well for (T78), assigning the inference a probability close to 1. To test that it works as expected for P2 of (T78), we substitute "Few musicians read music" for "Every musician reads music". BIS assigns the conclusion false with a high probability to H of (T78), which is a reasonable estimated value for this variant of the argument pattern.

3.2 Generalised quantifiers and generics

In addition to examples with universal quantification, our test suite includes cases with other generalised quantifiers (few, most, etc.), and generics expressed as bare plurals.

(T54) P1. Few people are basketball players.
P2. Basketball players are taller than most non basketball players.
P3. John is a basketball player.
H. John is taller than most people.
Label: COMPARATIVE ADJECTIVE, MODAL ADVERB

3.3 Gradation, Adjectives, and Comparatives

We test our system on gradation, adjectival modification, and comparatives against the sorts of examples discussed in the linguistic semantics literature, e.g. by Klein (1980).

In English, a (positive) adjective such as "tall" can be turned into the comparative "taller", which has the property of transitivity (if $X\,taller\,Y$ and $Y\,taller\,Z$, then $X\,taller\,Z$). Moreover, the adjective and the comparative derived from it are related in meaning, as the example (T15) illustrates (H holds given that P1 and P2 hold). If $X\,taller\,Y$ and $tall(Y)$, then $tall(X)$.

BIS does well on tasks where an inference involves transitivity of a relation expressed as a comparative adjective. The system computes a probability of 1 for (T15).

(T15) P1. Mary is tall.
P2. John is taller than Mary.
H. John is tall.
Label: COMPARATIVE ADJECTIVE, TRANSITIVITY

3.4 Modal Adverbs

BIS handles modal adverbs, such as "usually", "always", "rarely" (sometimes called adverbs of frequency), which can be used to turn categorical judgments into probabilistic ones. Examples where both adverbs of frequency and quantifiers (including generics and generalised quantifiers) interact are particularly interesting. They are not only complex from a computational modelling perspective. They are also semantically difficult. Their interpretations are not straightforward.

As we are interested in probabilistic judgments, we test our model against similar examples that contain probabilistic modifiers, as in (T17) below. The main difference between (T15) and (T17), is that (T17) contains the modal frequency adverbs "usually" and "always" that interact with a comparative. They trigger frequency based semantic relations that condition probabilistic inferences.

(T17) P1. John is always as punctual as Mary.
P2. Sam is usually more punctual than John.
H. Sam is more punctual than Mary.
Label: QUANTIFIER, MODAL/TEMPORAL ADVERB

BIS computes a value of 0.7 for this example.

3.5 Vagueness

Finally, we take up measure predicates that involve vagueness.

(T38) P1. Mary is 190 centimeters tall. Mary is tall.
P2. Molly is 184 centimeters tall. Molly is tall.
P3. Ruth is 180 centimeters tall. Ruth is tall.
P4. Helen is 178 centimeters tall. Helen is tall.
P5. Athena is 166 centimeters tall. Athena isn't tall.
P6. Artemis is 157 centimeters tall. Artemis isn't tall.
P7. Joanna is 160 centimeters tall. Joanna isn't tall.
P8. Kate is 162 centimeters tall. Kate isn't tall.
P9. Christine is x centimeters tall.
H. Chistine isn't tall
Label: QUANTIFIER, MODAL ADVERB

We perform 15 runs for 18 values of x uniformly distributed in the range 145cm to 201cm. BIS generates encouraging results for this case. The conclusion is always true with high probability where x is lower than 166, the highest measurement judged to be not tall. There is a slight deviation when x is 166, where 3 cases return false with

probabilities from 0.5 to 0.85. In the intermediate cases, for which we expect the vagueness effect, we see a near incremental increase of false judgments values with higher probabilities. From 174 cm and upwards, the system returns false consistently. After the lowest judgment of tallness (178), the system returns false with very high probability (many cases are 1).

BIS offers a gradient treatment of measure predicates (through tweaking of priors), expressing vagueness, but it is not yet fully incremental or stable. We seem to be on the right track in our treatment of measure predicates. Improving this aspect of BIS is one of our priorities for future work.

4 Related and Future Work

We do not have any baseline to compare our system to. The only implemented approach similar to ours is the one proposed by Goodman and Lassiter (2015); Lassiter and Goodman (2017). This system is not tested against a test suite. Furthermore, it is not designed to deal with the range of syntactic structures or complex inference patterns that BIS handles. Adapting the Goodman-Lassiter model to allow for such testing would require changes that undermine a comparison with BIS.

Goodman and Lassiter (2015); Lassiter and Goodman (2017) implement a probabilistic semantics in WebPPL. They regard the probability of a declarative sentence as the most highly valued interpretation that a hearer assigns to the utterance of a speaker in a specified context. On this approach, speakers express unambiguous meanings in specified contexts through their utterances, and hearers estimate the likelihood of distinct interpretations as corresponding to those that the speaker intends to convey. Their account requires the existence of a univocal, non-vague speakers meaning that hearers seek to identify by distributing probability among alternative readings. Goodman and Lassiter posit a boundary point parameter for graded modifiers, where the value of this parameter is determined in context. They adopt a classical Montagovian treatment of generalised quantifiers, and their framework has limited coverage of syntactic and semantic structures.

We take the probability value of a sentence to be the likelihood that a competent speaker would endorse an assertion, given specified premises. Predication is intrinsically vague, and we do not assume a sharply delimited reading for a predication that hearers attempt to converge on by estimating the probability of alternative readings. All predication consists in applying a classifier to new instances, on the basis of supervised training. BIS does not posit a contextually dependent cut off boundary for graded predicates or non-graded predicates. Instead, we adopt an integrated approach to both types of classifier on which a property term allows for vague borders. BIS applies a probabilistic treatment of generalised quantifiers, and it covers higher-order quantifiers like *most*.

The design of BIS is inspired by the Bayesian compositional semantic framework proposed by Bernardy et al. (2018). But BIS differs from this framework in a number of important respects. First, it has a comprehensive syntax-semantics interface through GF parsing. Second, it is intended to cover inference in a systematic way, including logically valid, as well as probabilistic arguments. Third, BIS has considerably wider coverage than the framework of Bernardy et al. (2018), and it is constructed in such a way as to permit straightforward extension to new types of sentence structure and inference patterns.

van Eijck and Lappin (2012) distribute probability values for natural language sentences over the set of possible worlds. The probability of a sentence is the sum of the probability values of the worlds in which it is true. If these worlds are understood as maximal consistent sets of propositions, as in classical theories of formal semantics, then it is unclear how they can be represented in a computationally tractable way.[3] Our system avoids these problems by sampling only the individuals and properties (vector dimensions) required to estimate the probability of a given set of statements.

Cooper et al. (2015) propose a compositional semantics within a probabilistic type theory (ProbTTR). They take the probability of a sentence to be a judgment on the likelihood that a given situation is of a particular type, specified in terms of ProbTTR. They do not offer an explicit treatment of vagueness or probabilistic inference. It is also not clear to what extent their type theory is required to achieve a viable compositional probabilistic semantics.

Emerson and Copestake (2017a,b) provide a

[3] Lappin (2015) discusses the complexity problems posed by the representation of complete worlds.

probabilistic model in order to identify 'features' of objects in terms of the properties that apply to those objects. They build their model as a graphical probabilistic model. They also interpret universal and existential quantifiers from a probabilistic perspective. "As are Bs" is represented as a conditional probability of B given A, for all elements of the space, which is equal to the sum (integral) over all elements. To compute it, they make use of the variational inference for graphical probabilistic models.

Pfeifer and colleagues (Pfeifer and Sanfilippo, 2018; Pfeifer, 2013; Gilio et al., 2015) study inference in a probabilistic setting by estimating the probability of the conclusion given the probabilities of the premises. They employ p-validity by Adams (1998). To be p-valid the uncertainty of a conclusion in an inference should not increase the cumulative uncertainties of its premises. Their approach differs from ours in several ways. The main one is that we build a model (using Bayesian updating of priors) where the premises hold, and then we observe how probable the hypothesis is in this model. By contrast, they provide an analytic estimation of the conclusion, given its premises. They require that certain properties on conditional probabilities hold. Conditional probabilities are primitives for modelling an implication ("if A then B"). This allows them to avoid problems when estimating $A \rightarrow B$ when A is false. In the current work, we take "if...then..." statements to be cases of material implication ($A \rightarrow B = \neg A \vee B$) instead of conditional probability.

In future work we will explore the interpretation of the "if...then..." construction as a conditional probability, and we may incorporate Pfeifer and colleagues' insights into our semantics.

We plan to extend the current test suite to examples that contain phenomena which are not yet represented there. This will allow us to increase both BIS' coverage and its power. We intend to organise the test suite in a more structured way, by introducing a more systematic and fine-grained classification of example types, and example complexity. To illustrate what we have in mind, imagine that BIS gives the correct result for test suite example n_1, but fails on n_2, where the two cases are labelled as of the same type, but n_2 is simpler than n_1. Given such a typology and complexity hierarchy over examples, it becomes easier to detect the source of peculiar behaviour in the system. We will also take into account that some examples might not make sense in a probabilistic setting. As Suppes (1966) remarks, statistical syllogisms require a specific formulation in order to be well posed as probabilistic problems. Building such a structured test suite is a challenging task.

In our model, we have sentence-level and predicate-level negation, which we refer to as *weak* and *strong* negation, respectively. In this way, we obtain a logic which deviates from both classical first-order and intuitionistic logic. We will explore the formal properties of this alternative logic, and we will consider the most efficient way of encoding it in BIS.

We observed in the previous section that chained transitive inferences become problematic for BIS in proportion to the length of the chain. This is due to errors that originate in Monte Carlo Methods of approximating integrals, which BIS uses to generate its models. We will experiment with an alternative approach that calculates the required integrals symbolically. On this strategy BIS would invoke Monte Carlo Methods only as a last resort, when symbolic computation is not feasible.

5 Conclusion

This paper describes BIS, an implemented Bayesian system for probabilistic inference. We have tested BIS on a test suite for probabilistic and classically valid arguments, which we have constructed for this task. While the test suite is still under development, it is, to our knowledge, unique in that its examples make probabilistic judgments based purely on the knowledge contained in the premises. The arguments do not require additional world knowledge beyond the information contributed by the premises to support their conclusions. We will reorganise and extend this suite to achieve a fine-grained, labelled typology for its examples.

While BIS follows the approach outlined by Bernardy et al. (2018) in many respects, it handles a wider and richer range of phenomena. In addition to providing a systematic Bayesian inference system, it offers a unique treatment of vagueness through a distinction between two types of negation, and an alternative procedure for computing interpretations for quantifiers.

We have also noted some of the limitations of the current system. Most of these are due to the fact that BIS does not yet encode certain linguis-

tic phenomena. Other problems arise because of BIS's current method for sampling and computing (numerical) approximations. We will address these issues in future work. BIS's current level of performance suggests that it can be scaled up to a wide coverage semantic system for probabilistic natural language inference.

In extending our test suite our primary objective is to provide a better platform for evaluating probabilistic semantic approaches. One way of obtaining more reliable gradient judgments of entailment is to submit inferences to crowd source assessment on a four or five point scale, and to map this scale into probability values (or ranges of values). The mean judgments of such an annotated suite would provide a gold standard for evaluating the performance of a probabilistic inference system. We are also interested in testing BIS on a standard dataset for logical inference, like the FraCaS test suite, that is annotated for categorical inference judgments. Success would consist in assigning high probability to yes cases, low probability to no cases, and intermediate values to unk instances. We could also crowd source the FraCas set, and use the mean judgments that we obtain as the target values for our system.

Acknowledgements

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg. We are grateful to our colleagues in CLASP for helpful discussion of some of the ideas presented here. We also thank three anonymous reviewers for their useful comments on an earlier draft of the paper.

References

Ernest Adams. 1998. *A Primer of Probability Logic*. Stanford: CSLI Publications.

Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, and Shalom Lappin. 2018. A compositional Bayesian semantics for natural language. In *Proceedings of the International Workshop on Language, Cognition and Computational Models, COLING 2018, Santa Fe, New Mexico*, pages 1–11.

Johannes Borgström, Andrew D. Gordon, Michael Greenberg, James Margetson, and Jurgen Van Gael. 2013. Measure transformer semantics for Bayesian machine learning. *Logical Methods in Computer Science*, 9:1–39.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Stephen P. Brooks. 1998. Markov chain monte carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):69–100.

R. Cooper, D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman. 1996. Using the framework. Technical report LRE 62-051r, The FraCaS consortium.

R. Cooper, S. Dobnik, S. Lappin, and S. Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology*, 10:1–43.

Ido Dagan, Bill Dolan, Bernado Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15:1–17.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Guy Emerson and Ann Copestake. 2017a. Semantic composition via probabilistic model theory. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Guy Emerson and Ann Copestake. 2017b. Variational inference for logical inference. *CoRR*, abs/1709.00224.

Angelo Gilio, Niki Pfeifer, and Giuseppe Sanfilippo. 2015. Transitive reasoning with imprecise probabilities. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 95–105, Cham. Springer International Publishing.

N. Goodman and D. Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin and C. Fox, editors, *The Handbook of Contemporary Semantic Theory, Second Edition*, pages 655–686. Wiley-Blackwell, Malden, Oxford.

N. Goodman, V. K. Mansinghka, D. Roy, K. Bonawitz, and J. Tenenbaum. 2008. Church: a language for generative models. In *Proceedings of the 24th Conference Uncertainty in Artificial Intelligence (UAI)*, pages 220–229.

Noah D Goodman and Andreas Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org. Accessed: 2018-12-10.

Ewan Klein. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4(1):1–45.

Shalom Lappin. 2015. Curry typing, polymorphism, and fine-grained intensionality. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory, Second Edition*, pages 408–428. Wiley-Blackwell, Malden, MA and Oxford.

Daniel Lassiter and Noah Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194:3801–3836.

Niki Pfeifer. 2013. The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning*, 19(3-4):329–345.

Niki Pfeifer and Giuseppe Sanfilippo. 2018. Probabilistic semantics for categorical syllogisms of figure II. In *Scalable Uncertainty Management*, pages 196–211. Springer International Publishing.

Aarne Ranta. 2004. Grammatical framework. *Journal of Functional Programming*, 14(2):145–189.

Gareth O. Roberts and Jeffrey S. Rosenthal. 2004. General state space markov chains and mcmc algorithms. *Probab. Surveys*, 1:20–71.

Patrick Suppes. 1966. Probabilistic inference and the concept of total evidence. In Jaakko Hintikka and Patrick Suppes, editors, *Aspects of Inductive Logic*, volume 43 of *Studies in Logic and the Foundations of Mathematics*, pages 49–65. Elsevier.

J. van Eijck and S. Lappin. 2012. Probabilistic semantics for natural language. In Z. Christoff, P. Galeazzi, N. Gierasimszuk, A. Marcoci, and S. Smets, editors, *Logic and Interactive Rationality (LIRA), Volume 2*. ILLC, University of Amsterdam.