Predicates as Boxes in Bayesian Semantics for Natural Language

Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin,
and **Aleksandre Maskharashvili**
Gothenburg University, Department of philosophy,
linguistics and theory of science,
Centre for linguistics and studies in probability

firstname.lastname@gu.se

Abstract

In this paper, we present a Bayesian approach to natural language semantics. Our main focus is on the inference task in an environment where judgments require probabilistic reasoning. We treat nouns, verbs, adjectives, etc. as unary predicates, and we model them as boxes in a bounded domain. We apply Bayesian learning to satisfy constraints expressed as premises. In this way we construct a model, by specifying boxes for the predicates. The probability of the hypothesis (the conclusion) is evaluated against the model that incorporates the premises as constraints.

1 Introduction

Goodman et al. (2008) interpret natural language expressions as probabilistic programs, which are evaluated through Markov chain Monte Carlo (MCMC) methods. This technique assigns meanings to various phenomena, including graded adjectives (Lassiter and Goodman, 2017). Bernardy et al. (2019, 2018) combine this approach with the idea (present in much recent computational linguistic literature (Mikolov et al., 2013, 2018; Pennington et al., 2014) (but which can be traced back to Gärdenfors (1990)) that individuals are encoded as points in a multidimensional space. Using this approach they construct Bayesian models of inference for natural language. While these models work well for many cases, they generate serious complexity problems for others.

In this paper we propose a simplified geometric model that allows us to reduce the need for sampling, and the complexity that it can create. In certain cases we eliminate sampling altogether. We model properties as (unions of) boxes, and we identify individuals as points. To estimate the probability of a predication being true, we determine the likelihood that an individual, a set of individuals, or another property is contained in a box corresponding to a predicate. This framework gives us a more tractable procedure for evaluating the probability of sentences exhibiting the same syntactic and semantic constructions that the approaches proposed by Bernardy et al. (2019, 2018) cover, but it extends to all representations of predicates in a probabilistic language.

The alternative system for evaluating arguments that we propose brings us closer to the prospect of a wide coverage probabilistic natural language inference system. Such a system will be useful for the Recognising Textual Entailment task (Dagan et al., 2009), which encompasses non-logical arguments based on real world knowledge and lexical semantics. It can also be applied in other NLP tasks that rely on probabilistic assessment of inference.

2 Interpretation of predicates as boxes

An underlying assumption of a Bayesian interpretation of natural language is that one has an (immanent) space of all (relevant) individuals, and predicates are represented as measurable subspaces of this space.

We treat every linguistic predicate as a box in an n-dimensional euclidean space. (A scaled n-cube whose faces are orthogonal to the axes.) To simplify computing the volume of a box, we also take the underlying space of individuals itself to be a box of a uniform density. Without loss of generality, we can assume that this box is of dimension 1 in all directions, and it is centred at the origin. We denote this unit box by U.

Formally, with each predicate P we associate two vectors of dimension n, P^c and P^d, where P^c is the centre of the box and P^d_i is the (positive) width of the box in dimension i. Hence, the subspace associated with P is the subspace $S(P)$

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 333–337
Turku, Finland, 30 September – 2 October, ©2019 Linköping University Electronic Press

given by

$$P(x) = \forall i. ||x_i - P^c_i|| < P^d_i$$

Note that $S(P)$ itself never extends past the complete space:

$$S(P) = U \cap \left\{ x \middle| \forall i. ||x_i - P^c_i|| < P^d_i \right\}$$

(A box could isomorphically be defined using lower and higher bounds P^l and P^h with $P^c = 0.5(P^h + P^l)$ and $P^d = 0.5(P^h - P^l)$).

Typically, P^c and P^d will be themselves sampled. In our experiments, P^c_i is taken in the uniform distribution on $[0, 1]$, while $1/P^d_i$ is taken in a beta distribution with parameters $a = 2, b = 8$.

2.1 Relative clauses

Boxes are closed under intersections. Thus if we use the expression $P \wedge Q$ to denote the intersection of the predicates P and Q, we have $(P \wedge Q)^l_i = \max(P^l_i, Q^l_i)$ and $(P \wedge Q)^h_i = \min(P^h_i, Q^h_i)$. The centre and the width of the box ($(P \wedge Q)^c$ and $(P \wedge Q)^d$ respectively) are recovered using the habitual formula.

2.2 Quantifiers

With this in place, we can interpret quantifiers. In classical formal semantics the phrase "every P is Q" is interpreted by

$$\forall x. P(x) \rightarrow Q(x)$$

A naive translation of this formula yields:

$$\forall x. (\forall i. ||x_i - P^c_i|| < P^d_i) \rightarrow (\forall i. ||x_i - Q^c_i|| < Q^d_i)$$

Enforcing this condition *as such* in a probabilistic programming language is expensive. It requires:

1. Sampling an individual x.
2. Verifying if x satisfies the hypothesis ($P(x)$). If not, go back to point 1.
3. Check if x satisfies the conclusion ($Q(x)$). If not, stop, otherwise loop back to point 1.

Typically this loop is iterated thousands of times, in order to ensure that we do not miss (too many) points x where P holds but Q does not. Even though optimisations are possible in the general case, the above algorithm is inefficient. The condition that it tests is really intended to check the inclusion of $S(Q)$ in $S(P)$. Because both spaces are boxes, this test can be done without sampling by checking the following geometric constraint:

$$\forall i. P^l_i \leq Q^l_i \wedge P^h_i \leq Q^h_i$$

where $P^l = P^c - P^d$ and $P^h = P^c + P^d$.

2.3 Generalised quantifiers

Generalised quantifiers can also be efficiently implemented with box models. Consider the phrase "most P are Q." Following Bernardy et al. (2019, 2018), "most P are Q" can be interpreted as

$$V(P \wedge Q) \geq \theta V(P)$$

for a suitable proportion θ matching the semantics of "most" in the context. Here, $V(P)$ stands for the measure of $S(P)$ in the space of individuals. In general, this measure is given by

$$V(P) = \int \mathbf{1}(P(x)) \, \mathsf{PDF}_{Ind}(x) \, dx$$

with $\mathbf{1}(c)$ being 1 if the condition c is true and 0 otherwise. Considering that individuals are elements in a high-dimensional space, if either the density of individuals PDF_{Ind} or $P(x)$ is non-trivial, the above integral is often non-computable symbolically. (This is the case, for example, if PDF_{Ind} is a Gaussian distribution). Instead it must be approximated numerically, often using a Monte Carlo method.

By contrast, if $S(P)$ is a box in a uniform space, then we have

$$V(P) = \prod_i (P \wedge U)^d_i$$

Thus, "most P are Q" is interpreted as follows:

$$\prod_i (P \wedge Q)^d_i \geq \theta \prod_i P^d_i$$

2.4 Graded predicates

We want our models to support predicates that correspond to comparative degree properties. To accommodate these properties we associate a degree function with predicates.

The degree to which an individual x satisfies a property P is

$$s(P, x) = 1 - \max \left\{ \frac{||x_i - P^c_i||}{P^d_i} \middle| i \in [1..n] \right\}$$

This definition entails that the subspace corresponding to a predicate coincides with the space where its degree of satisfaction is positive. Formally:

$$x \in S(P) \text{ iff. } s(P, x) > 0$$

Additionally, the maximal degree of satisfaction is 1.

The phrase "x is taller then y" is interpreted as x satisfying the $Tall$ predicate to a larger degree than y:

$$s(Tall, x) > s(Tall, y).$$

Predicates formed from positive comparatives are also boxes. For example, the predicate $P(x) = [\![$"x is taller than k"$]\!]$ for some constant individual k is a box centered at $Tall^c$ and whose widths is given by

$$Tall^d = (1 - s(Q, y))P^d$$

2.5 Negation and union

Boxes are closed under intersection, but not under negation nor union. Thus, in general, a predicate is represented by a union of disjoint boxes. If a predicate can be represented by a *single* box, we call it a *box-predicate*. Measuring the volume and checking intersection of general predicates is a straightforward combinatorial extension of the corresponding box-predicate algorithms.

However, general predicates cannot be associated with a degree, in the sense of the previous section – only box-predicates can. This limitation is in fact a welcome result. It correctly rules out phrases like "John is more not-tall than Mary" or "John is more tall or happy than Mary" as infelicitous, but sustains "John is shorter than Mary". Traditional formal semantic approaches to gradable predicates (e.g. Klein, 1980; Kennedy, 2007) have a problem excluding cases like "John is more not-tall than Mary."

3 Comparison with bisected multivariate Gaussian model

We highlight a few important differences between the present box model and the bisected multivariate Gaussian model proposed by Bernardy et al. (2018).

In the Gaussian model, individuals are represented as vectors, sampled in a multivariate Gaussian distribution of dimension k, with a zero mean and a unit covariance matrix. A (unary) linguistic predicate is represented as a pair of a bias b and a vector d: d is obtained by normalising a vector sampled in the same distribution as individuals, while b is sampled in a standard normal distribution. The interpretation of a predicate can be understood as a hyperplane orthogonal to d with b being the shortest distance from the origin to the hyperplane. An individual satisfies a predicate P if it lies on the far side of the hyperplane, as measured from the origin. Hence, every predicate partitions the vector space into two parts: one of individuals satisfying P, and one of individuals satisfying not-P.[1]

In the box model, a linguistic predicate is represented as a box, and individuals satisfy the predicate if they lie inside the boundary of the box. Here, individuals are sampled in a uniform distribution. For gradable predicates, we see here an important difference: in the Gaussian model, an individual has a higher degree of P if and only if it lies further from the origin, while in the box model, having a higher degree of P means lying closer to the center of the box.

Priors differ between the Gaussian model and the box model. In the Gaussian model, an arbitrary individual has a 0.5 chance of satisfying an arbitrary predicate when no additional information is given. In contrast, in the box model, the same situation has a 0.15 chance of holding. While these priors are somewhat arbitrarily chosen, they reflect the different geometric structures of the two models. If, in the box model, an arbitrary predicate corresponded to a box covering half the space, any additional predicate would force intuitively very non-probable configurations of the space. In particular, each additional predicate would have a lower probability of holding for an arbitrary individual.

The Gaussian model evaluates the size of a predicate by estimating the volume of the space beyond the corresponding hyperplane using MCMC sampling. Similarly, degrees of predicate inclusion (used for the interpretation of generalised quantifiers) are calculated by estimating the volume of the overlapping space. Approximation of the volumes by sampling is required since the density of individuals in the space is non-trivial. By contrast, in the box model, the volume of a predicate extension can be calculated by symbolic

[1] In a recent work, Bernardy et al. (2019) propose a Gaussian model in which a predicate divides the space into three disjoint sections, but we set aside a detailed comparison with that model.

means, since every such extension is a box, the surrounding space is bounded, and individuals are distributed uniformly in this space.

The evaluation of inclusion differs between the two models. In the Gaussian model, a predicate P is fully contained in a predicate Q if and only if the corresponding hyperplanes are parallel and the distance of P from the origin is greater than the distance of Q from the origin. This configuration is stochastically impossible to obtain, meaning that the system would fail to evaluate any argument with "every P is Q" among its premises. This condition can be relaxed in several ways to make it satisfiable. Bernardy et al. (2019, 2018) sample elements from P and check if all satisfy Q. The issue with this approach is that if the predicate P is far from the origin, then the density of individuals is so low that sampling does not converge in a reasonable time. Another possibility is to check that the angle between the planes defining P and Q is less than a certain threshold α. But this raises another issue: implication is no longer transitive (even if the angle between P and Q is less α and the angle between Q and R is also, it does not follow that P and R are also separated by an angle less than α.)

By contrast, the box model interprets inclusion of P in Q by placing the box for P strictly inside the boundaries of Q. This is easier to obtain, by sampling the dimensions for the box P within the box Q. As a consequence, any predicate contained in another predicate has a strictly lower chance of holding for an arbitrary individual than any arbitrary predicate has.

We did a preliminary evaluation of our model using the testsuite for probabilistic inference developed by Bernardy et al. (2019). While there is no gold standard to evaluate against, the results obtained by our model are more stable than the ones obtained from the Gaussian model. This is likely to depend on the indeterminacy introduced by sampling in the Gaussian model: increasing the number of samples would improve stability, but also lead to longer computation times.

4 Related Work

Boxes in Euclidean spaces are simple objects, and as such they have already been considered as geometric representations of predicates. Vilnis et al. (2018) use boxes to encode WordNet lexical entries (unary predicates) in order to predict hypernyms. Like us, they take the distribution in the vector space to be uniform, and the probability of a predicate is defined as the volume of the corresponding box. In our work, we use a Bayesian model. It is best suited to represent a small number of predicates, and to fully model the uncertainty of the boundary for each box. Vilnis et al. (2018) opt for a neural network to learn a large number of box positions. This is appropriate, given that their data set is the complete WordNet hypernym hierarchy. Their model converges on a single mapping of predicates to precise box boundaries, rather than to a distribution of such mappings.

We have not yet tested the box representation of words by Vilnis et al. (2018) for our task, but we plan to do so in future work. As our approach applies Bayesian sampling, we will need to modify the sizes of certain boxes to deal with a data set of this kind. It is important to recall that because their representations are learned for the purpose of detecting the WordNet hypernymy, they do not need to contain any additional lexical information not required for this task.

5 Future Work and Conclusion

We present an approach to natural language inference based on Bayesian probabilistic semantics for natural language. It differs from the work of Bernardy et al. (2019, 2018) in several respects. The main distinction is that we model predicates as boxes contained in a unit box, while they use (infinite) subsets of a vector space equipped with a Gaussian density. The density of the distribution in the current approach is uniform, which allows us to construct a more computationally efficient system for estimating the probability of the conclusion of an argument, given its premises. Our system is more stable than the one described by Bernardy et al. (2019) when tested against their test suite.

We have been relying on expert subjects for judgments on the strength of probabilistic inferences. In future work, we plan to collect crowdsourced data to ground these estimates or try to crowd source existing categorically annotated datasets like the FraCas test suite (Cooper et al., 1996), and use the mean judgments that we obtain as the target values for our system. Another way of testing our system would be to evaluate against the categorically annotated datasets, e.g. the FraCaS test suite. Success in this case would consist in assigning high probability to yes cases, low probability to no cases, and intermediate values to unknown instances.

Instead of boxes, one could use arbitrary convex polytopes. This would give a more precise, but more computationally expensive model. We leave further evaluation of this trade-off to future work.

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

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.

Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, and Aleksandre Maskharashvili. 2019. Bayesian inference semantics: A modelling system and a test suite. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM), Minneapolis*, pages 263–272. Association for Computational Linguistics.

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. `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

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

Peter Gärdenfors. 1990. Induction, conceptual spaces and AI. *Philosophy of Science*, 57(1):78–95.

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.

Christopher Kennedy. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1):1–45.

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

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

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272. Association for Computational Linguistics.