Classification systems: Combining taxonomical and perceptual lexical meaning

Bill Noble **Staffan Larsson** **Robin Cooper**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{bill.noble@, sl@ling, cooper@ling}.gu.se

Abstract

Lexical meaning includes both perceptual and logical aspects. We present a method for combining a taxonomy with perceptual classifiers, and show that in the few-shot setting, it outperforms other methods of injecting taxonomical information in image classification. We use this method to define witness conditions for types in a rich type system with probabilistic type judgments and suggest how such a type system can be used as the basis for a new type of hybrid NLU architecture.

For words like *red*, *apple*, and *hug*, part of what it means for a person—or indeed an artificial NLU system—to understand the word's meaning is the ability to recognize that some object is red, or an apple, or that some event is one in which hugging is taking place. Marconi (1997) calls this **referential competence**. Another mode of understanding is supported by **inferential competence**, which has to do with the relationship that certain lexical items have with one another—a system that infers that John is not married from the sentence *John is a bachelor* demonstrates inferential competence with the words *bachelor* and *married*. Marconi (1997) argues that neither of these competencies are reducible to the other, meaning that a comprehensive theory of lexical meaning must explain both referential and inferential ability.

In this paper, we propose a framework for combining taxonomical information, which supports an inferential competence, with perceptual classifiers, which implement referential competence. This *classification system* is formalized in a rich type theory with probabilistic type judgments, meaning it can be integrated in a formal semantics based on Type Theory with Records (Cooper et al., 2015).[1]

1 Classifier-based perceptual meaning

While distributional methods of representing meaning have achieved a lot of success, many have argued that that relying on exclusively *ungrounded* meaning representations has fundamental limitations (Harnad, 1990; Bender and Koller, 2020; Bisk et al., 2020).

Classifier semantics offers a way to ground lexical meaning, operating on the intuition that part of what it means to understand the meaning of a word is to be able to identify instances of it based on perceptual input.

In one approach to classifier semantics (e.g., Schlangen et al., 2016; Silberer et al., 2017), the parameters of a learned classifier (for example, the relevant row of a liner classifier's weight matrix) are regarded as a distributed representation of the meaning of the word. Alternatively, it is possible to regard the classifier itself, as a function of type $f : \mathit{PerceptualData} \rightarrow [0, 1]$, that provides the semantics of the relevant word (e.g., Larsson, 2020a). Here, both the parameters of the classifier and the classification algorithm are considered to be part of the perceptual meaning, whereas in the distributed approach, the classification algorithm is simply a means by which a distributed representation is learned.

In this work, we take a functional approach to classifier semantics. Because they can (at least for one-place predicates) be considered analogous to Montague's $e \rightarrow t$ type, it is natural to integrate classifiers-as-functions in a type-theoretic approach to compositional meaning. Furthermore, classifiers have the nice theoretical property that they can distinguish between intentional identity and extensional equivalence (Muskens, 2005; Lappin, 2012; Larsson, 2020b).

A *multi-class classifier*, C for a set of labels, L is a function that takes an input and produces a prediction among the labels in the form of a prob-

[1] A PyTTR implementation of a classification systems based on convolutional visual classifiers is available online here: https://github.com/GU-CLASP/classification-systems. We also make available the code for the experiments conducted in Section 4.

ability distribution. We will consider multi-class classifiers that take perceptual data as input:[2]

$$C : \mathit{PerceptualData} \rightarrow (L \rightarrow [0,1]),$$

subject to the restriction that for any input a, $\sum_{l \in L} C(a)(l) = 1$.

2 Folk taxonomies

A *folk taxonomy* is a hierarchically structured collection of conceptual categories that is *common ground*, in the sense of Clark (1996), in a certain speech community. We wish to invoke a more general notion than that of scientific or technical taxonomies that rely on an authoritative reference for their common ground status. Folk taxonomies by contrast can be informal, emerging from the communicative needs of a particular community and changing in response to changes in the environment. Such a taxonomy can also be established in an *ad hoc* way between a group of speakers, grounded in a particular interaction.

For now, we define a taxonomy in set theoretic terms. A taxonomy takes the form

$$\mathrm{Tax} := \langle \mathrm{Taxon}, \mathrm{Set}(\mathrm{Set}(\mathrm{Tax})) \rangle,$$

where Taxon is the label for a taxonomical category. A taxonomy bottoms out in pairs of the form $\langle \mathrm{Taxon}, \varnothing \rangle$, which we refer to as *leaf taxons*.

Notice that the second element of Tax is a set of *sets* of taxonomies. To see why this is, we will first introduce the notion of a *distinction*, which is a pair that takes the following form:

$$\mathrm{Dist} : \langle \mathrm{Taxon}, \mathrm{Set}(\mathrm{Taxon}) \rangle.$$

Consider this taxonomy:

$$\begin{array}{l}\langle \mathit{object}, \{\{\langle \mathit{animal}, \{\\ \{\langle \mathit{mammal}, \{...\}\rangle, ..., \langle \mathit{bird}, \{...\}\rangle\},\\ \{\langle \mathit{herbivore}, \varnothing\rangle, \langle \mathit{omnivore}, \varnothing\rangle, \langle \mathit{carnivore}, \varnothing\rangle\}\},\\ \langle \mathit{vegetable}, \{...\}\rangle, \langle \mathit{mineral}, \{...\}\rangle\}\}\rangle\end{array}$$

Here, animal is subject to two distinctions: the distinction based on diet, and the one that categorizes animals as mammals, birds, and so on.

In the following, we let $\mathit{dist} : \mathrm{Tax} \rightarrow \mathrm{Set}(\mathrm{Dist})$ be the function from a taxonomy to its distinctions.

A *genus-species* relation holds between a taxon and the (first component of) an element of one of its distinctions. In the above example, both mammal and $\mathit{herbivore}$ are species of animal.[3] Conceptually, the key feature of a distinction is that it implies an exhaustive partition of the genus into a set of mutually exclusive species. Note however that we need not assume every species is associated with a lexical item—there can, for example, be a catch-all species in cases where the named alternatives don't cover the entire genus.[4]

This leaves us with two main desiderata for when we start giving content to our taxonomy in the next section.

1. An instance of a species is an instance of its corresponding genus.

2. An instance of a genus is an instance of exactly one species in each of its distinctions.

3 Classification systems

By associating a word with a prediction class of a classifier, a system can be endowed with at least some referential competence. Similarly, associating a word with taxon gives a system some inferential competence in relation to other words embedded in the taxonomy. In this section, we describe a *classification system*, which combines classifiers and a taxonomy to integrate these two kinds of competence.

With this in mind, we will formalize a classification system as a rich Martin-Löf (1984)-style type system that allows for probabilistic type judgments (as in Cooper et al., 2015). Furthermore, we will assume that we can provide basic types with *witness conditions* that ground type judgments. From the perspective of an agent, a type's witness conditions are the methods by which an agent may judge something to be of that type (Cooper, forthc).

[2] In the remainder of the paper, we restrict our attention to classifiers and taxonomies of individuals, so we assume that *PerceptualData* is of a kind that corresponds to entities of type *Ind*. In general, however, we can also classify other kinds of entities (events, relations between individuals, etc.).

[3] For word senses, this is referred to as a *hypernym-hyponym* relation.

[4] Generally we would expect a conventionalized taxonomy to make distinctions in a systematic way; that is, where the species within a distinction are differentiated along some common dimension or set of dimensions. This intuition can be traced back at least to Aristotle's *Categories*. However, his is not a formal requirement of a taxonomy at this stage and nor could it be since, taxons are not yet associated with any kind of content that could be considered as features or establish differentia. Such content will come by way of classifiers in Section 3.

Suppose we have a taxonomy $\mathbf{T}$, and a classifier, C_d, for each distinction $d \in \mathit{dist}(\mathbf{T})$. For each taxon, t, in the taxonomy, we want to define a type, T_t, with the appropriate witness conditions such that $p(a : T_t)$ estimates the probability that a belongs to the taxon, according to the classifiers.

Intuitively, the classifiers give content to the distinctions of the taxonomy by *distinguishing* between species. The classifier is thus premised on the assumption that the object of classification certainly belongs to *one* the species, s_i, among which it distinguishes, meaning that it must in turn belong to the associated genus, g. In practice, this means that the classifier for a given distinction is trained on the subset of labeled data from the associated genus. The classifier's prediction, $C_d(a)(s_i)$, can thus be interpreted as the conditional probability that a has belongs to s_i, given that it belongs to g.

There is one taxon in the taxonmy—the root taxon—that is not a species in any distinction. Let T_{t^*}, which we will refer to as the *domain* classification system, be the type associated with the root taxon. We will assume that T_{t^*} is *universal* in the sense that it is witnessed by any object:[5]

$$p(a : T_{t_*}) = 1 \tag{1}$$

Every other taxon is a species in some distinction, meaning that we have a classifier associated with it. Let $d = \langle g, \{s_1, ..., s_n\}\rangle \in \mathit{dists}(\mathbf{T})$ be a distinction. We define auxiliary types, $T'_{s_1}...T'_{s_n}$ with witness conditions as follows:

$$p(a : T'_{s_i}) = C_d(a)(s_i). \tag{2}$$

That is, an object a is judged to be of type T'_{s_i} with probability equal to the probability assigned by the classifier for the corresponding distinction.

The interpretation of the classifier as providing a conditional probability suggests that we should define T_{s_i} such that:[6]

$$p(a : T'_{s_i}) = p(a : T_{s_i} \mid a : T_g) \tag{3}$$

We also want T_{s_i} to satisfy the desiderata from the end of Section 2, which can be restated as follows:

$$p(a : T_{s_i}) \leq p(a : T_g) \tag{4}$$

and

$$p(a : T_{s_i} \mid T_g) = 1 - \sum_{j \neq i} p(a : T_{s_j} \mid a : T_g) \tag{5}$$

With this in mind, we let the witness conditions for T_{s_i} be defined as the product of the probability assigned to T'_s and T_g:[7]

$$p(a : T_{s_i}) = p(a : T'_{s_i}) \cdot p(a : T_g) \tag{6}$$

By induction on the taxonomy and the base case of T_{t*}, this gives us well-defined witness conditions for for every taxon t.

Briefly, we will show that this definition meets each of our desiderata. In the following, let $\langle g, \{s_1, ..., s_n\}\rangle$ be a distinction. Without loss of generality, we consider the case of T_{s_i}.

We get (4) directly from (6), since $0 \leq p(a : T'_{s_i}) \leq 1$. As a result of (4) we may write $T_{s_i} \sqsubseteq T_g$—i.e., that T_{s_i} is a *subtype* of T_g (Cooper et al., 2015). Furthermore, this has the consequence that

$$p(a : T_g | a : T_{s_i}) = 1 \tag{7}$$

From Bayes Theorem and (7), we can prove (3):

$$\begin{aligned}
&p(a : T_{s_i} \mid a : T_g) \\
&\qquad = \frac{p(a : T_g \mid a : T_{s_i}) \cdot p(a : T_{s_i})}{p(a : T_g)} \\
&\qquad = \frac{p(a : T_{s_i})}{p(a : T_g)} \\
&\qquad = \frac{p(a : T'_{s_i} \cdot p(a : T_g)}{p(a : T_g)} \\
&\qquad = p(a : T'_{s_i})
\end{aligned}$$

Finally, (5) follows from (3) and the fact that $\sum_{i \leq n} C_d(a)(s_i) = 1$.

4 Empirical comparison

To investigate how well the classification system performs in practice, we compare it with two other plausible methods of combining classification with taxonomical hierarchy. We put aside type theory

[5] This assumption is convenient for simplicity, but it also works if T_{t_*} is given some constant prior or well-defined witness conditions as part of some larger type system in which the classification system is embedded.

[6] This corresponds to the probability that a is of type T_{s_i} given that it is of type T_g, though other notions of conditional judgments are possible in probabilistic type theory. See Larsson and Cooper (2021).

[7] Note that T_{s_i} has different witness conditions from that of the meet type $T'_{s_i} \wedge T_g$, as defined in Cooper et al. (2015), since the witness condition for the meet type is defined by the classical Kolmogorov (1950) equation for conjunction:

$$p(a : T'_{s_i} \wedge T_g) = p(a : T'_{s_i}) \cdot p(a : T_g \mid a : T'_{s_i}),$$

which is different since we can't assume that $C_d[s_i]$ is probabilistically independent from $C_{d'}[g]$, where d' is the distinction of which g is a species.

	Precision	Recall	F1
per-distribution	**0.93**	**0.90**	**0.90**
marginalization	0.90	0.86	0.82
hierarchy-agnostic	0.80	0.84	0.81

Table 1: Macro-averaged precision, recall, and F1 score for the three methods of incorporating hierarchy in classification.

for the moment and make a comparison based on metrics that are traditionally used for machine learning classification.

Dhall et al. (2020), proposes several possible methods of incorporating hierarchical information, including the *hierarchy agnostic* and *marginalization* methods that we compare against.[8]

The **hierarchy agnostic** method is the simplest and most common way of dealing with a taxonomically organized label set. Every label is considered by a single *multi-label* classifier, without respect to taxonomical hierarchy. There is thus no guarantee that the predicted probabilities will be consistent—the probability assigned to a genus label could be lower than the probability assigned to one of its species, for example. Hopefully the hierarchical relations inherent in the data encourages the classifier to learn a function that approximates the taxonomy.

In the **marginalization** method, a *bottom-up* classifier, is trained on the leaf nodes in the taxonomy. Labels at higher levels are predicted by marginalizing the leaf node probabilities—the probability of a genus label is computed as the sum of the probability of its species labels. Note that this method assumes that the leaf labels are disjoint, meaning that it only works for taxonomies in which there is on distinction per genus.

The system described in Section 3 is will be referred to as the **per-distinction** method. As described there, we train a classifier for each distinction and compute the probability of a given label as the product of the classifier output and the probability assigned to its parent label.

We test each method on a simple synthetic dataset shapes with different colors and sizes. The data was generated with a hierarchical stochastic process reflected in the taxonomy of the labels given to each item. Images were encoded with a convolutional autoencoder, which was pre-trained on images from a larger unstructured sample space.

Each method used simple single-layer linear classifiers trained by stochastic gradient descend through backpropagation. The marginalization and per-distinction classifiers use softmax activations with categorical cross-entropy as the loss function, and the hierarchy agnostic classifier uses a sigmoid activation and binary cross entropy with the indicator function of the item's actual label set. Table 1 gives a summary of the results of the classifiers in a few-shot classification scenario with 5 training instances and 100 testing instances for each leaf label. A separate set of 100 development items were used to choose the best model after 10 epochs of training. For the precision, recall and F1 metrics, the predicted classes were chosen in a greedy fashion from the top of the taxonomy, taking the label with the highest probability consistent with the label chosen at the previous level.

Consistent with Dhall et al. (2020), we find that both methods that explicitly take the label hierarchy into account out-perform the hierarchy agnostic method. In the few-shot experiment reported here, our per-distribution method performed best, though we note that this advantage is less pronounced with more training examples.

5 Conclusion

In this paper we have focused on the problem of integrating perceptual and logical meaning on a lexical level. To do this, we have embedded perceptual classifiers as witness conditions for types in a type system that respects a taxonomical structure. Our method for doing this is based on the intuition that such a taxonomy gives rise to a collection of *distinctions*, whose content can be defined by multiclass classifiers. We have compared our method of embedding classifiers at each node in a taxonomy to other strategies for classifying in a taxonomically structured label space suggested by (Dhall et al., 2020). Future work should also consider the possibility of learning the label hierarchy on the fly, as Bengio et al. (2010) does. Embedding such a hierarchy in a type system may present additional challenges, but allowing for changes to the taxonomy would be necessary to full model the plasticity of the lexical semantic structures used by natural language speakers.

[8] Dhall et al. (2020) also tests a *per-level* and *masked per-level* method, which are arguably most similar to what we propose here. We do not reproduce those tests because marginalization tended to out-perform them in Dhall et al. (2020)'s experiments. Like marginalization, the per-level and masked per-level methods assume that there is a single distinction per genus.

We have also left open the looming question of compositional semantics. We presented classification systems as a rich type system in order to suggest a way forward in this regard. Our proposal is compatible with Type Theory with Records (TTR), which can be used to define version of compositional semantics (Cooper et al., 2015). Indeed, TTR has been used for compositional semantics with perceptual classifier-based meaning (Larsson, 2013, 2017). The issue remains, however of how to compose the types we define in Section 3.

Composing classifiers-as-functions is no easy task.[9] For a given object a, one can compute the probability that a witnesses both T_1 and T_2 simply by taking judgments for T_1 and T_2 separately. The difficulty comes when one needs to reason hypothetically, as is necessary in NLI. What is the likelihood that *some* object of type T_1 is also of type T_2? One way forward is to find a way to compose the classifiers for T_1 and T_2 directly, as Monroe et al. (2017) does for color terms. Another option is to use the classifiers to to sample from conditioned space of objects. Something like this is the basis of the system proposed by Bernardy et al. (2019), though it is not perceptually grounded. In order for that to work, the embedding space of $\mathit{PerceptualData}$ would have to be regularized in such a way that admits sampling, which could potentially be achieved by using a variational autoencoder (Kingma and Welling, 2014).

Aside from compositionality, there remain many questions on the side of lexical representation, such as that of polysemy. It would seem that certain words may appear in multiple places in a taxonomy. The meaning of a word may be ambiguous among a *set* of such corresponding types. So far we have only discussed predicative nouns. Adjectives, and verbs, including transitive verbs admit a similar treatment, but that leaves quantifiers and function words, among others.

Finally, we only discuss perceptual and taxonomical aspects of meaning, but there are other aspects of meaning, including other inferential aspects. How would we represent, for example, that *being from the Champagne region* is an aspect of the meaning of *champagne* (the beverage)? In Marconi (1997)'s schema, this fact would be treated as an aspect of inferential competence. Certainly we should not expect the inference to be derivative of a perceptual classifier for champagne, but it does not fit neatly as taxonomical information either. A more sophisticated type system is needed to incorporate lexical information of this kind.

[9] Importantly, it is a different task from the one of composing distributed representations learned through classification. See Moro et al. (2019) for more on that task.

Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *ACL 2020*.

Samy Bengio, Jason Weston, and David Grangier. 2010. Label Embedding Trees for Large Multi-Class Tasks. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

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 2019)*, pages 263–272, Minneapolis, Minnesota. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Robin Cooper. forthc. *From Perception to Communication: A Theory of Types for Action and Meaning*. Oxford University Press.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic Type Theory and Natural Language Semantics. In *Linguistic Issues in Language Technology, Volume 10, 2015*. CSLI Publications.

Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. 2020. Hierarchical Image Classification using Entailment Cone Embeddings. *arXiv:2004.03459 [cs, stat]*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Conference Proceedings: Papers Accepted to the International Conference on Learning Representations (ICLR)*, Calgary.

A. N. Kolmogorov. 1950. *Foundations of the Theory of Probability*. New York: Chelsea Pub. Co.

Shalom Lappin. 2012. An Operational Approach to Fine-Grained Intensionality. *UCLA Working Papers in Linguistics, Theories of Everything*, 17:180–186.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Staffan Larsson. 2017. Compositionality for perceptual classification. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short Papers*.

Staffan Larsson. 2020a. Discrete and Probabilistic Classifier-based Semantics. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 62–68, Gothenburg. Association for Computational Linguistics.

Staffan Larsson. 2020b. Extensions are Indeterminate if Intensions are Classifiers. In *SemDial 2020 (WatchDial) Workshop on the Semantics and Pragmatics of Dialogue*, page 10, Waltham, MA and online.

Staffan Larsson and Robin Cooper. 2021. Bayesian Classification and Inference in a Probabilistic Type Theory with Records. In *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, pages 51–59, Groningen, the Netherlands (online). Association for Computational Linguistics.

Diego Marconi. 1997. *Lexical Competence*. Language, Speech, and Communication. MIT Press, Cambridge, Mass.

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

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Daniele Moro, Stacy Black, and Casey Kennington. 2019. Composing and Embedding the Words-as-Classifiers Model of Grounded Semantics. *arXiv:1911.03283 [cs]*.

Reinhard Muskens. 2005. Sense and the Computation of Reference. *Linguistics and Philosophy*, 28(4):473–504.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1213–1223, Berlin, Germany. Association for Computational Linguistics.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2017. Visually Grounded Meaning Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2284–2297.