Language Tasks and Language Games: A Critical Review of Methodology in Current Natural Language Processing Research

David Schlangen
Computational Linguistics / Department of Linguistics
University of Potsdam, Germany
DRAFT (V0.1) – PLEASE DO NOT CITE – *comments welcome*
schlangen@uni-potsdam.de

Abstract

"This paper introduces a new task and a new dataset", "we improve the state of the art in X by Y" – it is rare to find a new natural language processing paper (or AI paper more generally) that does *not* contain such statements. What is mostly left implicit, however, is the assumption that this necessarily constitutes progress, and what it constitutes progress towards. Here, we formalise the normally impressionistically used notions of *language task* and *language game* and ask under which conditions success on such tasks and games might translate into progress on the goal of modelling general language competence. We observe that the approach of modelling constrained tasks requires making a *separability hypothesis* about language comptence as a whole, and an *exhaustivity hypothesis* about the sub-competences modelled, and that for these, one might profitably turn to linguistics and cognitive science, to a greater degree than currently done.

1 Introduction

Recently, seemingly ever other natural language processing paper introduces a new task and a new dataset.[1] We join some other recent papers (Yogatama et al., 2019) in asking whether there is any coherence to this approach, under which conditions this could be considered progress, and towards what. What we do differently, however, is to look at the fundamental assumptions behind this approach. We formalise central notions, in order to make different attempts comparable.

In our argumentation, we distinguish between *language tasks*, such as for example "describe this image", or "translate this sentence"—that is, single-step tasks that involve in an essential way natural language material, but not necessarily *only* language material—; *micro worlds*, which are environments that produce disinterested responses to actions, thereby possibly simulating the behaviour of independently existing systems; these environments can then enable *dialogue games*, as repeated and connected language tasks. We define these notions first and think about general ways of evaluating their relevance. This will give us the vocabulary to discuss, in the second part of the paper, some current tasks and games in more detail, analysing the motivations for their creations given in the original papers.

2 Tasks, Worlds, and Games

We formalise these notions as follows.

2.1 Definitions

2.1.1 Tasks

Definition 1 *A* Language Task *is a tuple* $(S, A, \mathcal{L}, D_T)$*, where:*

- S *is a (possibly infinite) set of* states*,*
- A *is a (possibly infinite) set of* actions*,*
- *with either the states in* S *or the actions in* A *(or both) having as part* natural language expressions*, and*
- $\mathcal{L} : S \rightarrow A$ *is a function that maps a state* $s \in S$ *to an action* $a \in A$*, where*
- *the mapping* $\mathcal{L}$ *conforms to* task description D_T*.*

This very general definition essentially covers much, if not all of natural language processing. For example, translation can be seen as a language task where the state space consists of expressions in one language, the action space of expressions in another language, and the task description, to

[1] Not quite, but not very far. Looking at the 2018 long and short paper proceedings of ACL and EMNLP, we get 94 hits for "introduce new dataset", 20 hits for "introduce new corpus", and 101 hits for "introduce new task".

which the mapping must conform, is that in each pair in the mapping, the second element be a translation of the first. (What exactly it means for a mapping to "conform to" a task description is left unspecified here. That there does not seem to be a formal and precise way to define this notion is one of the problems that this paper aims to discuss.)

Here, we are interested in tasks that more directly exhibit some form of *understanding* of language than translation does—where understanding clearly also is involved, but is more latent. We call such tasks *Language Understanding Tasks*; what exactly makes a language task an understanding task is a topic of this paper. To give an example of an understanding task that we will discuss in more detail later, we can define *image description* as the task of mapping a state consisting of an image and a sentence to a judgement action (from the space $\{true, false\}$), where the task description is to determine whether the sentence is true of the (situation depicted by the) image. It hence involves understanding of the description—as well as understanding of the image in order to apply the description to it.

We further talk of an *interpretation task* when the state involves a natural language expression, and a *generation task* when the action involves language. (A given task can be both.) Orthogonal to this, we can distinguish *reference tasks*, where non-linguistic and linguistic material are to be related, and *inference tasks*, where linguistic material is to be related.

A *task description* can be given informally, as in the examples above, making reference to theoretical or pre-theoretical constructs external to the definition, such as "translation" or "is true of". We call this an *intensional description*. Often, a task is also specified *extensionally* through the provision of a *dataset* of examples of the mapping (that is, pairs of state and action), $\mathcal{X} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where the assumption is that $(x, y) \in \mathcal{X} \rightarrow y = \mathcal{L}(x)$.

2.1.2 Worlds

Definition 2 *A* Micro-World *or* Environment *is a tuple* $(S, A, \mathcal{E}, R, D_W)$*, where function* $\mathcal{E} : S \times A \rightarrow S \times R$ *maps an action* a*, taken in state* s*, to a state* s' *and a reward* r*, and the mapping conforms to the* world description D_W.

The language competence of humans plays out in *repeated* task (which we will define below), not single-step ones as described above, and it plays out in contexts where language use is embedded in a non-linguistic context. To study such repeated, situated games, much recent work has made use of environment simulators that compute reactions to actions performed within them, in accordance to the assumed (or actual) rules of the domain they represent. We formalise such environments as mappings from an action to an environmental response (a state), where again the mapping conforms to a description of what it is intended to model.[2]

2.2 Games

We approach the definition of a *language game* via the more general notion of *interaction game*:[3]

Definition 3 *An* Interaction Game *is a tuple* (P, A, o, T, E, D_G)*, where:*

- $P = \{p_1, \ldots, p_n, N\}$ *is the set of* n *regular players* p*, together with one additional player* N *(for* Nature*).*
  N has a special status in that it does not have a strategic interest in the outcome of the game.

- $A = \{A_1, \ldots, A_n, A_N\}$ *is the set of action spaces, with one space per player.*
  Action types can be complex: e.g., (nav, s), for "navigation action, *south*", or $(utt, \text{"I don't know"})$ for "utterance of *I don't know*". If defined in the right way (for example by a recursive grammar), the set of action types can be infinite. Players chose actions from their space of available actions; the resulting action tokens a_j are

[2] In this desire to model the relevant aspects of a domain, and in the assumption that from dealing with a simulated environment transferable knowledge about dealing with the original enviroment can be achieved, this approach is reminiscent of the AI microworlds of the 1970s—"we see solving a problem often as getting to know one's way around a 'micro-world' in which the problem exists." (Minsky and Papert, 1972)—and perhaps, as we will discuss, susceptible to similar kinds of critiques as these attempts (Dreyfus, 1981; Marr, 1982).

[3] *Pace* Wittgenstein (1953/84), we get to define what a game is. Or we could go with (Suits, 1978), who is happy to define games as rule-guided activities of voluntary attempt to overcome unnecessary obstacles. His concepts of *prelusory goal*, which can be stated independently of the game (e.g., in football (soccer), "make the ball be in the opponent team's goal"); *constitutive rules*, which make reaching that goal more difficult than necessary (e.g., by disallowing to just grab the ball and carry it to the goal); and *lusory attitude*, which is to accept the complications posed by the constitutive rules, can inform the design of games that work via crowdsourcing.

associated with their originator through a function $\mathsf{a} : A_i \to P$, and with a position in the sequence of actions that have been performed since the beginning of the interaction through a function $\mathsf{t} : A_i \to \mathbb{N}$.

- $o : P_i \times A_i \to \mathcal{P}(P)$ *(for* $P_i \in P, A_i \in A$*) is the* observability function *that specifies which types of actions by which player can be observed by which subset of the players.*
  In normal cases, one would assume that players can observe their own actions, and that Nature observes all actions; but this allows for the specification of deviant cases.

- $T : \emptyset \cup (P \times A) \to \mathcal{P}(P)$ *is the* turn taking rule *that specifies who can act next, depending on who did what last. It also specifies who can start the game.*
  In a *free initiative* setting, any player can act at any time; in a strict turn based setting, the current player would always be excluded from the set of next players.

- $E : S \to V$ *is the* evaluation rule *that maps a sequence of action tokens* $\langle a_1, \ldots, a_m \rangle$ *into an* evaluation*, where the set of possible evaluations* V *includes at least one positive one (e.g.,* success*) and one negative one (e.g.,* failure*).*
  The evaluation is made known to the players when a positive or negative outcome has been reached. If it is not, or if it is does not contain outcomes denoted as positive or negative, we call the resulting structure an interaction *setting*, rather than an interaction game.

- D_G *finally is the* game description *which specifies which, if any, otherwise existing activity the game is meant to approximate.*

The well-known *Gridworld* game (see e.g. Sutton and Barto (1998)) for example can be represented in these terms as being an interaction game with one regular player (the agent) interacting with Nature, $P = \{p_1, N\}$. The agent can only perform navigation actions: $A_1 = \{(nav, \mathrm{n}), \ldots\}$, Nature informs on the resulting available navigation options, $A_N = \{(\mathit{inform}, (n, w)), \ldots\}$, with the information that Nature relays coming from a microworld that simulates the grid and the movement on it.

Gridworld does not involve language, and hence is not an example of a *language* interaction game. As an example of a language interaction setting that is not a game, we can define *free chat interaction* in our terms as involving two players and an inert Nature that does not intervene: $P = \{p_1, p_2, N\}$, $A_1 = A_2 = \{(utt, \alpha)\}$, $A_N = \emptyset$, T is a constant function into P (free initiative), all actions are observed by all, E is a constant function into $\{\mathit{undecided}\}$.

2.3 What makes a good task, world, game?

2.3.1 Tasks

As we have seen above, tasks are often exemplified by the provision of a dataset of examples of the task being executed. Evaluating such datasets is relatively straightforward. First, a dataset should be *verified*, which is to check whether the provided input / output pairs can indeed be judged correct relative to the task (in its intensional description). If the examples are collected specifically for the purpose of exemplifying the task, this is the process of controlling annotation, and standard methodologies exists (Artstein and Poesio, 2008).

Validating a dataset is a less formalised process. It comprises arguing that the dataset exemplifies the task intension well. E.g., pairs only of images of giraffes and sentences describing them may *not* be seen as exemplifying the general task of image description well, while perhaps exemplifying the task of *giraffe image description*.

Another way to evaluate datasets is by providing a model of the task learned on parts of it, and testing it on the other part (for which a comparison, or "loss" function on A must be provided as well). If a model can "solve" the dataset even when not given information that for theoretical or pre-theoretical reasons is seen to be crucial, the dataset can be considered an unsatisfactory exemplification of the task. E.g., in a *visual question answering* setting, if in a dataset all and only the expressions that mention giraffes are true, a model would not need to take the images into account at all to perform well, which would be evidence that the dataset is deficient relative to the task description.

How can the task itself be evaluated? This is easy, if it has a direct value to a consumer (such as translation presumably has), which can be measured. If the consumer is a computer system that processes the output of the task further, the burden of evaluation is simply shifted to the system

as a whole. If the interest is in replicating with a theoretically motivated model performance characteristics of humans attempting the task, the task can be evaluated for its power helping distinguish between different modelling choices.

A recent trend, however, has been to motivate tasks in a different way, neither via their inherent practical use, nor as answering questions about language processing as implemented in humans. The argument roughly goes as follows: To be good at task T, an agent must possess a set C_T of capabilities (of representational or computational nature). If the $c \in C_T$ are capabilities that competent language users can be shown or argued to possess—let's call the set of these capabilities of a competent language user C_L, so that $C_T \subseteq C_L$—then being able to model these capabilities (via modelling the task) results in progress towards the ultimate goal, which is to model competent language use.

This argument is at best incomplete, of course. Unless it is claimed that $C_T = C_L$ — which is one way to understand Turing's proposed task (Turing, 1950) — the progress claim requires an additional *separability assumption* according to which no interactions are to be expected between capabilities in C_T and those in $C_L \setminus C_T$. Otherwise, it is a possibility that a model of T (and with it, of C_T) overfits and makes choices incompatible with other capabilities. Put differently, without that assumption, the task can only serve as a *negative* benchmark, in that it is *necessary* to be able to handle it, and failing it means failing in approaching "general AI"; succeeding in it alone would not be any indication of whether a model that can be made sufficient has been found.

Further, with regards to c, an *exhaustivity assumption* must be made, which says that T brings out all there is to c, and that another task T', if it requires c as well, could be handled by a model of c built with only T in mind. Or else, we only have "c as required by T", which does not help us indicate progress beyond T.

This points at a weekness, perhaps, which is that if you don't have a theory of how C_L decomposes, neither of these assumptions can really be backed up very well. In any case, this discussion shows us how one can attack a proposed task: By claiming that it does not in fact require the assumed capabilities and can be solved without possessing them (as in the dataset example above); by arguing that the required capabilities are not interestingly involved in language competence; or by claiming that for language competence, the covered capabilities are not separable from such that are not. Only the first of the three lines of attack involves methods and knowledge that is not specific to the domain of (human) language processing.

2.3.2 Worlds

How do you evaluate a micro world? By whether it makes the right abstractions – and enables interesting tasks. Making the right abstractions means that those aspects of its real counterpart that matter are modelled, as specified in the (informal) *world description*. (It is *valid*, relative to the description.) Some environments can be modelled exhaustively (e.g., the rules of games like Chess or Go, or Settlers of Catan). Others can not (e.g., "the interior of a house, through which an agent moves, from the perspective of that agent").

Trying to make progress through modelling tasks in simulated worlds entails making another separability hypothesis, which assumes that the natural competence of handling the world as a whole is separable into handling various parts of it, which can be "knit" together to form the whole.[4]

(Adams et al., 2012) propose the following characteristics of "artificial general intelligence" environments: "C1. The environment is complex, with diverse, interacting and richly structured objects. C2. The environment is dynamic and open. C3. Task-relevant regularities exist at multiple time scales. C4. Other agents impact performance. C5. Tasks can be complex, diverse and novel. C6. Interactions between agent, environment and tasks are complex and limited. C7. Computational resources of the agent are limited. C8. Agent existence is long-term and continual."

(Baroni et al., 2017b)

Robotics is a field that had to learn painfully how hard it is to transfer methods from simulation to the real world.

(Brooks, 1991b): "In order to really test ideas of intelligence it is important to build complete

[4] "[W]e feel [the micro-worlds] are so important that we plan to assign a large portion of our effort to developing a collection of these micro-worlds and finding how to embed their suggestive and predictive powers in larger systems without being misled by their incompatibility with literal truth. We see this problem—of using schematic heuristic knowledge—as a central problem in Artificial Intelligence. [...] In order to study such problems, we would like to have collections of knowledge for several 'micro-worlds', ultimately to learn how to knit them together." (Minsky and Papert, 1972).

agents which operate in dynamic environments using real sensors."

(Brooks, 1991a) about connectionists "building" systems: "in simulation only–no connectionist has ever driven a real robot in a real environment, no matter how simple)"

2.3.3 Games

How do you evaluate a language game? For the kinds of interactions that it produces, and whether they reflect phenomena that you are interested in.

Whether it makes the embedded language tasks more interesting – that is, whether it increases the range of required capabilities in the desired way.

Cite Sellars on Language Games?

Somewhere there must be room for a *linguistic* evaluation... Is this really dialogue? (Whatever it means to be really dialogue..)

Games produce data of a different kind... Games are endlessly reproducible... Interactions..... Number of possible chess games. Number of possible language games, if the goal is sufficiently complex and the action space is infinite (as it would be, if freely formulated utterances are allowed)? Any given data set will always underrepresent the space... Model is an agent that can participate in the game, and can be tested against humans.. How is that different from testing a translation model on more data?

Distribution from which the data is drawn is more varied and less likely to be described by surface regularities?

The results of games – the trace of the game playing – are, except in the most trivial cases, long sequences of actions, most of which are *re*actions. In less rigid games, and especially in language games, typically there is a large variability in how similar effects can be achieved. All this leads to data sparsity, and hence the problem that a dataset represents the game in which it was created less well than a dataset can represent a task.

The model becomes an agent.. needs more state...

Any kind of dialogue would be a language game in that sense. Booking a table at a restaurant via the phone. The agent that you infer from data from that task needs to have all the capabilities that are exhibited by the real person that has produced data (of the side that is being modelled).

But what about artifical games? Why do we set them up? Because we want to focus on certain capabilities and not be distracted by those others. So, again, a separability assumption, according to which the included capabilities are not catastrophically hurt by being separated from the non-included.

Perhaps no wonder that the games that are currently popular are just very slight extensions of existing tasks. Visual dialogue: visual question answering + a little bit of co-reference resolution. (Where visual question answering already is set up in such a way as to be a re-formulation of computing denotations, as answering the questions does not require any reasoning about why it was asked.)

Experience: Games make it harder to control which set of capabilities will be required to model how humans approach the game, unless the action space is strongly restricted (and other constitutive factors such as the turn-taking rules and the observability function are as well).

3 Some Example Tasks

3.1 Visual Question Answering

Task Definition Described in our terms (and not as in the original papers, which will be referenced below), *visual question answering* (VQA) is a language understanding task,[5] where a state consisting of an image (or set of images) I together with a natural language question Q is to be mapped to an "action" a, which here can be seen as an *utterance action* that is directly specified by the uttered expression α.

Formally, the task then is $(S, A, \mathcal{L}, D_T)$, with $S = \mathcal{I} \times \mathcal{Q}$ (where $\mathcal{I}$ is the set of all images, and $\mathcal{Q}$ the set of all questions; presumably restricted to one particular natural language); $A = \mathcal{A}$, the set of all potential answers (again, in one particular natural language); $\mathcal{L}$ is the mapping that conforms to the description D_T that $\mathcal{L}(\rho)$ be a correct answer to ρ, given image I.

The vocabulary of formal semantics allows us to make explicit the intuitive requirement that is hidden in the term "correct answer", namely that the answer be *true* of the image, and *relevant* to or *resolving* of the question.[6] For simplicity, we look at *polar* questions first.

Let $\ulcorner \cdot \urcorner$ be a function that takes an expression and returns an abstract representation of its

[5] More specifically, it is an *interpretation task*, as it involves language on the input side; if the question is a *wh*-question, it also a *generation task*.

[6] See (Ginzburg, 1995) for the notion of *resolving* a question.

meaning-relevant elements and structure; $[\![\cdot]\!]$ is a function that takes such a representation and builds up, in a compositional way out of its components, another function. This function takes as argument normally a mathematical structure, the *model*, that is meant to represent the "world" against which the expression is to be evaluated. Following (Schlangen et al., 2016; Schlangen, 2019), we can assume that the image can directly serve as model, so that $[\![\ulcorner \rho' \urcorner]\!]^I$ is the answer to the question, given the image (with ρ' being the assertive positive version of the question ρ, e.g., "it is raining" from "is it raining?"). The mapping $\mathcal{L}$ can hence be made precise as being a function that maps ρ to (a verbalisation of) $[\![\rho']\!]$. Non-polar questions simply require an additional operation which combines an assertion out of question and answer (e.g., "which sport is being played?"—"Tennis." to "Tennis is being played"), so that the task is to find an expression α such that $[\![apply(\ulcorner \alpha \urcorner, \ulcorner \rho \urcorner)]\!]^I = 1$.

Figure 1 shows two examples of such image and expression pairs (here, the expression already is in assertive form, but the underlying task is one of polar VQA), from (Suhr et al., 2018).

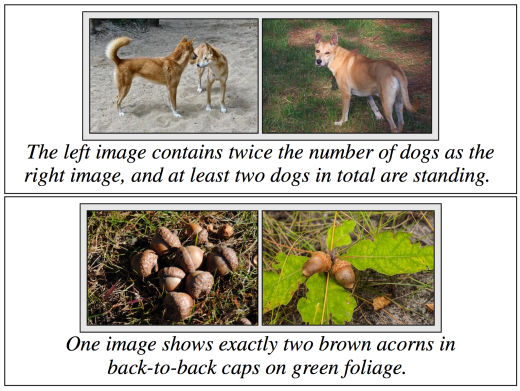

Figure 1: NLVR^2 Examples (from (Suhr et al., 2018))

Before we look at the motivation for studying this task given in the literature, we briefly note that the task perhaps is slightly more complicated than our formalisation reveals. The translation function $\ulcorner . \urcorner$ might not have a unique value for a given input, and so a decision has to be made on how the evaluation should proceed. From inspection of the datasets (described below), it appears that something akin to a *principle of charity* (Wilson, 1958) has been applied, according to which the best effort is to be made to make the resulting assertion true, if possible.

Motivation The original paper by Antol et al. (2015) that introduced the task and the first large-scale dataset provides a veritable smorgasbord of motivations, which are worth citing in full and relating to the discussion from above:[7]

- *"[There is] a belief that multi-discipline tasks like image captioning are a step towards solving AI."*

  The goal is "solving AI", and this goal can be approached in steps.

- *"What makes for a compelling AI-complete task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require multi-modal knowledge beyond a single sub-domain (such as CV) and (ii) have a well-defined quantitative evaluation metric to track progress."*

  The task of visual question answering is "AI-complete"—which is typically used to mean that it comprises all aspects of general intelligence, and solving it is equivalent to solving general AI—as it requires multi-modal knowledge.[8]

- *"Open-ended questions require a potentially vast set of AI capabilities to answer – fine-grained recognition (e.g., What kind of cheese is on the pizza?), object detection (e.g., How many bikes are there?), activity recognition (e.g., Is this man crying?), knowledge base reasoning (e.g., Is this a vegetarian pizza?), and commonsense reasoning (e.g., Does this person have 20/20 vision?, Is this person expecting company?)."*

  In particular, it involves these capabilities, which then appear to be necessary (or, if the term "AI-complete" is to be taken seriously, sufficient) for solving AI.

- *"This goal-driven task is applicable to scenarios encountered when visually-impaired*

[7] Our intention is not to single out one particular paper for claims made that go beyond the actual modelling contribution made in it. Our intention is to move the discussion of how a task fits into the larger context, at least in this paper, from the introduction section to the main body of a paper.

[8] This seems to illustrate an interesting semantic change in the use of this label. In Eric S. Raymonds "hacker jargon file", version 2.8.2, 1991, the implication is that a problem labeled as such is "just too hard" [http://catb.org/esr/jargon/oldversions/jarg282.txt].

users or intelligence analysts actively elicit visual infor- mation."

The task is also intrinsically valuable for certain user groups. (E.g., see (Gurari et al., 2018), which also shows that the kinds of questions asked by people interested in their answer is quite different from that in the popular VQA datasets.)

Datasets As mentioned above, Antol et al. (2015) introduced the first large scale dataset for this task. However, as became clear quickly, and noted *inter alia* by Jabri et al. (2016), this dataset could be handled competitively by models that were deprived of information information that according to the informal task description should be considered crucial: In many cases, the question alone makes the answer predictable, without inspection of the image, due to biases in the underlying image corpus (where the most frequent type of sports is tennis) or in the way questions were generated (existential polar questions seem to have been triggered by the actual presence of objects, and hence were predominantly to be answered in the positive). The criticism then is of the *validity* of the dataset, and not of the task itself. This kind of criticism was then addressed by Goyal et al. (2017) with the construction of less biased (and hence more valid) corpus.

A different line of work on this task was motivated by a perceived shortcoming in the set of capabilities engaged by the existing datasets. (Andreas et al., 2016) note that "questions in most existing natural image datasets are quite simple, for the most part requiring that only one or two pieces of information be extracted from an image in order to answer it successfully", which is to be understood as challenging "[c]ompositionality, and the corresponding ability to answer questions with arbitrarily complex structure" not enough. To that end, they introduce the SHAPES dataset which pairs synthetic images with synthetic, programmatically generated sentences that contain spatial relations.

Johnson et al. (2017) followed the same intuition and provided another dataset of synthetic images paired with programmatically generated questions. They further explain how the dataset challenges compositionality, as it's handling requires "[understanding of] unseen combinations of objects and attributes", and tests "visual reasoning abilities such as counting, comparing, logical reasoning, and storing information in memory". Along the same lines, the *Cornell Natural Language Visual Reasoning* corpus (NLVR, Suhr et al. (2017)) pairs synthetic images with natural sentences (collected via crowd sourcing) that contain spatial relations and quantification. NLVR^2 (Suhr et al., 2018) finally, from which the example above was taken, finally pairs natural images with natural sentences.

Using the terminology introduced above, we can see this line of work as pointing out a need for the function $[\![\cdot]\!]$ to be set up (or learned) in such a way that for an expression α composed out of expressions β, γ, the computation of $[\![\ulcorner\alpha\urcorner]\!]$ should involve $[\![\ulcorner\beta\urcorner]\!]$ and $[\![\ulcorner\gamma\urcorner]\!]$, in some way – or at least, performance should be commensurate to such a composition taking place.[9] (E.g., when a red house and a green hut were seen in training, a green house or a red hut should be recognised during testing.)

3.2 Natural Language Inference

Task Definition *Natural language inference* (NLI) is a *language interpretation task* where a state consisting of a pair of expressions (α, β) is to be mapped to a decision that categorises the type of entailment relation that holds between the elements of the pair, typically into one of ENTAILMENT, NEUTRAL, CONTRADICTION. Often, α is called the *premise*, and β the *hypothesis*.

Formally: $(S, A, \mathcal{L}, D_T)$, with $S = \mathcal{E} \times \mathcal{E}$, where $\mathcal{E}$ is the set of expressions (presumably from a given natural language), $A = \mathcal{R} = \{\text{ENTAILMENT}, \text{NEUTRAL}, \text{CONTRADICTION}\}$, $\mathcal{L}$ is the mapping that conforms to the task description D_T that $\mathcal{L}(\alpha, \beta) = R$, where $R \in \mathcal{R}$ and it is the case that $R(\alpha, \beta)$.

In formal semantics, the relation ENTAILMENT (from which the others can be derived) has a clear interpretation—in fact, it has *two* clear interpretations, that are equivalent in formal systems that have the property of soundness and completeness: As a syntactic notion, *entailment* relates to provability (and would typically be written $\vdash$). As a semantic notion, it is an inclusion relation between sets of models, stating that there can't be a model that makes the premise true, but not the hypothesis (written $\models$). The relation that is intended in this task, as example **ADD** illustrates, is not

[9] See (Andreas, 2019) for a recent attempt to make explicit what the demand for "compositionality" entails, and how to evaluate representations for how well they meet it.

quite this, as it may require arguably non-semantic common sense or typicality knowledge. It is more one that belongs to non-monotonic logic (Brewka et al., 1997), and should be interpreted as "if this is the case, it is *typically* also the case".

Motivation Has long tradition (under the name of "recognising textual entailment"), it has seen renewed recent interest due to the availability of larger resources...

"The ability to recognize such semantic relations is clearly not a *sufficient* criterion for language understanding: there is more to language understanding than just being able to tell that one sentence follows from another. But we would argue that it is a minimal, *necessary* criterion." (Condoravdi et al., 2003)

"The recognition of textual entailment is without doubt one of the ultimate challenges for any NLP system: if it is able to do so with reasonable accuracy, it is clearly an indication that it has some thorough understanding of how language works. Indeed, recognising entailment bears similarities to Turings famous test to assess whether machines can think, as access to different sources of knowledge and the ability to draw inferences seem to be among the primary ingredients for an intelligent system. Moreover, many NLP tasks have strong links to entailment: in summarisation, a summary should be entailed by the text; paraphrases can be seen as mutual entailment between T and H; in IE, the extracted information should also be entailed by the text." (Dagan et al., 2006).

"The task of natural language inference (NLI) is well positioned to serve as a benchmark task for research on NLU.", "In particular, a model must handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lexical and syntactic ambiguity" (Williams et al., 2018)

Is this contradictory? In logic, the interest in entailment is that it is a formal relation that doesn't need semantics. Here, it is set up as something that is necessary for language understanding, but also requires common sense knowledge... So it is not a formal relation.. Is it about language, or about the world?

For pre-training. E.g., BERT? Already in original paper.

A model that is good at NLI has nothing to say about VQA.

Discuss notion of a *benchmark*. Needs idea of necessary and separable capability. Hints at multi-task setups? What if benchmark becomes target?

Datasets Original (large-scale) dataset: (Bowman et al., 2015)

"SNLI falls short of providing a sufficient testing ground for machine learning models in two ways. First, the sentences in SNLI are derived from only a single text genre—image captions—and are thus limited to descriptions of concrete visual scenes, rendering the hypothesis sentences used to describe these scenes short and simple, and rendering many important phenomena—like temporal reasoning (e.g., yesterday), belief (e.g., know), and modality (e.g., should)—rare enough to be irrelevant to task performance. Second, because of these issues, SNLI is not sufficiently demanding to serve as an effective benchmark for NLU, with the best current model performance falling within a few percentage points of human accuracy and limited room left for fine-grained comparisons between strong models." "additionally interested in constructing a corpus that facilitates work on domain adaptation and cross-domain transfer learning" (Williams et al., 2018)

4 Some Example Environments

Baroni overview article.. ParlAI... Malmo...

Look at how they are motivated.. Or use overview article as starting point?

Habitat.

(Savva et al., 2019; Adams et al., 2012; Johnson et al., 2016; Urbanek et al., 2019; Baroni et al., 2017a; Xia et al., 2018; Yan et al., 2018; Misra et al., 2018; Côté et al., 2018; Bennett and Shatkhin, 2018; Anderson et al., 2018; Savva et al., 2017; Gordon et al., 2017; Brodeur et al., 2017; Chang et al., 2017; Janarthanam and Lemon, 2011; Baroni et al., 2017b; Byron et al., 2007; Yamauchi et al., 2013)

Note on use of "embodiment". There are reasons to suspect that the proponents of *embodied cognition* (Wilson and Foglia, 2017) would be particularly impressed by the use of this term for agents operating in simulated environments, at least as they are currently set up. A central claim in that field is that much apparent computation is done by the body in interaction with the environment. Most of the models that are presented that operate within these environments are only embodied insofar as they have a viewpoint, but other-

wise are massless idealised points of central cognition.

4.1 Vision Environments

Manipulable and non-manipulable. All these house simulators, and street view simulators.

4.2 Text-Only Environments

TextWorld, ParlAI / bAbI

5 Some Example Games

Reference games.

Navigation game.

Interactive / Embodied Question Answering.

Visual Dialogue.

Agreement Games.

6 Transfer Learning and Multi-Task Learning

Analyse this in terms of the framework developed here. Is this really a way beyond the separability hypothesis?

Multi-task: is that defining T that comprises T', T''? Not quite.. It is typically set up as a game that switches between Ts.. Rather than combining the sets of capabilities... Argue that dialogue is the way forward, as it combines all capabilities... Towards the Turing test....

(Wang et al., 2019) GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

Also mention Mikolov Roadmap towards machine intelligence?

7 Conclusions

Conclusion: Go deeper, and go wider. Deeper: Connect to cognitive science (not neuroscience), to back up how you dissociate competences. Wider: Model dialogue, which is naturally and parallely multi-task.

... and ideally this would be shown with a nice analysis of a dialogue example...

Sequentially multi-task vs. parallely multi-task... Dialogue is the way forward... But then, what about methodology? Wouldn't it be nice if all these things were nicely separable? You train a model first on this, then on that, then on that... Fine-tuning one each task, then turning to the next...

Dialogue isn't really a language task, per se... It's the $C_T = C_L$ task, as Turing knew...

Conclusion: re-kindle connection to cognitive science? Ideas about computational-level modelling... Connect to *psychology*, rather than neuroscience.

Need theory of C_L, and how it decomposes, to avoid the fate of the drunk in the well-known joke, who searches for the lost key where the light is and not where it most likely was lost.

Remember Marr's level of analysis, especially the computational level (Marr, 1982). What is being computed, and – crucially – why? What purpose does it serve in the cognition of the system as a whole.

References

Sam Adams, Itmar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, J. Storrs Hall, Alexei Samsonovich, Matthias Scheutz, Matthew Schlesinger, Stuart C. Shapiro, and John Sowa. 2012. Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine*, 33(1):25–42.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR 2018*.

Jacob Andreas. 2019. Measuring Compositionality in Representation Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–15.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural Module Networks. In *Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition (CVPR 2016)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistic*, 34(4):555–596.

Marco Baroni, Armand Joulin, Allan Jabri, Germàn Kruszewski, Angeliki Lazaridou, Klemen Simonic, and Tomas Mikolov. 2017a. CommAI: Evaluating the first steps towards a useful general AI. *arXiv*, pages 1–9.

Marco Baroni, Claes Strannegård, David L. Dowe, Katja Hofmann, Kristinn R. Thórisson, Jordi Bieger, Nader Chmait, Fernando Martínez-Plumed, and José Hérnandez-Orallo. 2017b. A New AI Evaluation

Cosmos: Ready to Play the Game? *AI Magazine*, 38(3):66.

Andrew Bennett and Max Shatkhin. 2018. Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction. In *EMNLP 2018*, pages 2667–2678.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Gerhard Brewka, Jürgen Dix, and Kurt Konolige. 1997. *Nonmonotonic Reasoning: An Overview*. Number 73 in CSLI Lecture Notes. CSLI Publications, Stanford, CA, USA.

Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. 2017. HoME: a Household Multimodal Environment. *ArXiv*.

Rodney A Brooks. 1991a. Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Rodney A. Brooks. 1991b. The Role of Learning in Autonomous Robots. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory (COLT '91)*, pages 5–10.

Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating Instructions in Virtual Environments (GIVE): A Challenge and an Evaluation Testbed for NLG. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

Angel Chang, Angela Dai, Thomas Funkhouser, Manolis Savva, and Shuran Song. 2017. Matterport3D : Learning from RGB-D Data in Indoor Environments. *ArXiv*.

Cleo Condoravdi, Richard Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G Bobrow. 2003. Entailment, intensionality and text understanding. In *Proc. of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. TextWorld: A Learning Environment for Text-based Games. *ArXiv*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Hubert L. Dreyfus. 1981. From micro-worlds to knowledge: AI at an impasse. In John Haugeland, editor, *Mind Design*. MIT Press.

Jonathan Ginzburg. 1995. Resolving questions I. *Linguistics and Philosophy*, 18:459–527.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2017. IQA: Visual Question Answering in Interactive Environments. *ArXiv*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR 2017*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *CVPR*.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting Visual Question Answering Baselines. In *European Conference on Computer Vision (ECCV)*.

Srini Janarthanam and Oliver Lemon. 2011. The GRUVE Challenge : Generating Routes under Uncertainty in Virtual Environments. In *ENLG '11 Proceedings of the 13th European Workshop on Natural Language Generation*, pages 208–211.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR 2017*, pages 1988—-1997.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. *IJCAI International Joint Conference on Artificial Intelligence*, 2016-Janua:4246–4247.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco, USA.

Marvin Minsky and Seymour Papert. 1972. Progress Report on Artificial intelligence. Technical report, MIT Artificial Intelligence Laboratory, Cambridge, Mass., USA.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction. *ArXiv*.

Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. 2017. MINOS: Multimodal Indoor Simulator for Navigation in Complex Environments. *ArXiv*, pages 1–14.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. *ArXiv*.

David Schlangen. 2019. Natural language semantics with pictures: Some language & vision datasets and potential uses for computational semantics. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Gothenburg.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*, Berlin, Germany.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A Corpus of Natural Language for Visual Reasoning. In *Proceedings of the 2017 meeting of the Association for Computational Linguistics (ACL 2017)*.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A Corpus for Reasoning About Natural Language Grounded in Photographs. In *Proceedings of NIPS 2018*, Montreal, Canada.

Bernard Suits. 1978. *The Grasshopper: Games, Life, and Utopia*. The University of Toronto Press, Toronto, Canada.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning*. MIT Press, Cambridge, USA.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to Speak and Act in a Fantasy Text Adventure Game. *ArXiv*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR 2019*, pages 1–20.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

N L Wilson. 1958. Substances without Substrata. *Review of Metaphysics*, 12(4):521–539.

Robert A. Wilson and Lucia Foglia. 2017. Embodied cognition. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2017 edition. Metaphysics Research Lab, Stanford University.

Ludwig Wittgenstein. 1953/84. *Tractatus Logicus Philosophicus und Philosophische Untersuchungen*, volume 1 of *Werkausgabe*. Suhrkamp, Frankfurt am Main.

Fei Xia, Amir Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson Env: Real-World Perception for Embodied Agents. In *CVPR 2018*.

Takashi Yamauchi, Mikio Nakano, and Kotaro Funakoshi. 2013. A Robotic Agent in a Virtual Environment that Performs Situated Incremental Understanding of Navigational Utterances. In *SIGdial 2013*, August, pages 369–371.

Claudia Yan, Dipendra Misra, Andrew Bennett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. 2018. CHALET : Cornell House Agent Learning Environment. *ArXiv*.

Dani Yogatama, Cyprien de Masson D'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and Evaluating General Linguistic Intelligence. *ArXiv*, pages 1–14.