Reading Between the Lines*

Loizos Michael
Department of Computer Science
University of Cyprus
loizosm@cs.ucy.ac.cy

Abstract

Reading involves, among others, identifying what is *implied* but not expressed in text. This task, known as textual entailment, offers a natural abstraction for many NLP tasks, and has been recognized as a central tool for the new area of Machine Reading.

Important in the study of textual entailment is making precise the sense in which something is implied by text. The operational definition often employed is a subjective one: something is implied if humans are more likely to believe it given the truth of the text, than otherwise. In this work we propose a natural objective definition for textual entailment.

Our approach is to view text as a partial depiction of some underlying hidden reality. Reality is mapped into text through a possibly stochastic process, the author of the text. Textual entailment is then formalized as the task of accurately, in a defined sense, recovering information about this hidden reality.

We show how existing machine learning work can be applied to this information recovery setting, and discuss the implications for the construction of machines that *autonomously* engage in textual entailment. We then investigate the role of using multiple inference rules for this task. We establish that such rules cannot be learned and applied in parallel, but that layered learning and reasoning are necessary.

1 Introduction

Text understanding has long been considered one of the central aspects of intelligent behavior, and one that has received a lot of attention within the Artificial Intelligence community. Many aspects of this problem have been considered and extensively studied, and frameworks have been developed for tasks such as summarization, question answering, syntactic and semantic tagging. The importance of text understanding has greatly increased over the past few years, following the recognition that the web offers an abundant source of human knowledge encoded in text, on which machines can capitalize (see, e.g., Reading the Web [Mitchell, 2005]). It has also been suggested that a robust and viable way for machines to acquire commonsense knowledge, similar to that employed by humans, is through learning from natural language text (see, e.g., Knowledge Infusion [Valiant, 2006]). A new area has, in fact, emerged with the goal of extracting knowledge from text, dubbed Machine Reading [Etzioni *et al.*, 2006].

Traditional Natural Language Processing tasks and techniques are useful components of this ambitious goal. Yet, the emphasis shifts from extracting knowledge *encoded within* a piece of text, to that of understanding what text *implies*, even if not explicitly stated. As an example, consider the following sentence: "Alice held a barbecue party last weekend.". Traditional NLP tasks include recognizing the entities, tagging words with their part of speech, creating the syntactic tree, identifying the verbs and their arguments, and so on. Beyond these tasks, however, one may also ask what can be *inferred* from this sentence. Although this question might not admit a unique answer, a possible inference might be that the weather was good last weekend. In fact, the author of the sentence may be aware, or even take for granted, that readers will make such an inference, and she may choose not to explicitly include this information. If machines are to understand the intended meaning of text, they should be able to draw similar inferences as those (expected to be) drawn by human readers.

The inference task can be seen as one of deciding whether the truth of a statement follows from the truth of some piece of text *and* some background knowledge. This task, known as *textual entailment*, has recently received considerable attention, since it naturally generalizes and abstracts many of the traditional NLP tasks [Dagan *et al.*, 2005]. Amongst the most successful approaches for this task is one that employs knowledge induced from a large corpus [Hickl *et al.*, 2006].

The ultimate goal, of course, is to have machines that *completely autonomously* acquire relevant background knowledge and subsequently use it to recognize textual entailment. Designing and implementing such machines would arguably be a concrete step forward in endowing machines with the ability to understand text by drawing those commonsense inferences that humans do when reading text. For this to happen, a crisp definition of textual entailment is first needed. The classical definition follows the semantics of logical implication (i.e., a possible worlds interpretation of entailment). This definition is, though, too rigid to be useful in practice. Instead, a more applied (operational) definition of textual entailment is used: "[a piece of text] entails [a statement] if the meaning of [the statement] can be inferred from the mean-

*This work was supported in part by grant NSF-CCF-04-27129.

ing of [the text], as would typically be interpreted by people" [Dagan *et al.*, 2005]. This definition requires, however, a subjective human gold standard, making the task an *inherently supervised* one, clashing with the goal for autonomy.

In this work we propose a definition for textual entailment that retains the formal objective aspects of the classical definition, yet it embraces the statistical nature of the operational definition, and avoids rigidness. Our approach is to view text as a partial depiction of some underlying hidden reality. This reality is mapped into a piece of text through the author of the text. Textual entailment is then formalized as the task of accurately, in a defined sense, recovering information about this underlying reality. An agent engaged in the textual entailment task is first given access to a corpus of text relevant to some domain of interest, and only then it is asked to recognize textual entailment. This training phase aims, from a semantics point of view, to make precise the type of statistical accuracy guarantees that are expected on the textual entailment task, and, from a practical point of view, to provide the agent with a means of acquiring relevant background knowledge.

We argue that for learning to be meaningful in this setting, a learner needs to be able to deal with missing information in its learning examples, both during the training phase, and the actual deployment phase, while making minimal assumptions on the nature of this missing information. We briefly review the key features of a learning framework called *autodidactic* [Michael, 2007], which respects these requirements, and show how it can be applied to the textual entailment task.

We continue to investigate the role of learning and reasoning with multiple rules in the textual entailment task. It is shown that chaining the conclusions of rules is provably beneficial. It is also established that these rules cannot be learned in parallel, but need to be learned iteratively, a layer at a time. These results we view as a formal prescription of how machines built for the textual entailment task should operate.

2 Text as Appearance of Some Hidden Reality

We propose that text be viewed as an appearance of some underlying and hidden reality. This view is naturally exemplified through the following scenario. Consider the case of a newspaper reporter present by chance at the scene of a bank robbery at the time of the event. The news reporter observes the event unfolding, and later documents her experience in a newspaper article. The article is subsequently read by the readers of the newspaper. From a reader's point of view, the reality of interest is the event of the actual bank robbery that took place. This reality, however, is not directly visible to the reader. The reader's only source of information about the event is the reporter's article, a piece of text. In this sense, this piece of text acts as an appearance, a partial observation, of the actual reality. Depending on the reporter, the article might contain many details of the actual event, or mention only the important, in the reporter's opinion, pieces of the event. The reporter might have a page limit to uphold in writing her article, and may be forced to drop even important aspects of the event. Even more so, the reporter might be biased in favor of or against the bank robber, and the text might describe the event noisily or inaccurately. Yet, from this piece of text the reader hopes to reconstruct (to some extent) the actual event.

To capture the nature of text as an appearance of some underlying reality, a means is required to model reality, text, and the author that ties the two together. We consider a fixed set $\mathcal{A} = \{x_1, \ldots, x_{|\mathcal{A}|}\}$ of attributes, and think of each attribute in $\mathcal{A}$ as an indicator variable of some aspect of the reality of interest. Certain domains may require representations that comprise a set of objects, and relations that hold amongst them. It has been shown, however, that *in the context of learning*, certain restricted forms of relational representations are not harder to learn than propositional representations [Roth and Yih, 2001; Valiant, 2000]. For this reason, and in the interest of simplicity, we restrict our attention to propositional representations. An event is represented as a binary string $\texttt{evn}$ of length $|\mathcal{A}|$, with the i-th bit of $\texttt{evn}$, denoted $\texttt{evn}[i]$, corresponding to the value of attribute x_i. The bank robbery event of our example could, for instance, correspond to the binary string $\texttt{evn} = 010001110111001$, where x_1 might stand for whether the bank robber was holding a gun (in this case "no"), x_2 for whether the bank robbery was successful (in this case "yes"), and so on. Some other bank robbery event would be represented by a different binary string, keeping, though, the same interpretation of the attributes.

By analogy, text describing an event is also represented as a string. To capture the fact that text is only partially depicting the event, we introduce a third value $\star$, which stands for "don't know". A piece of text describing the bank robbery event of our example could, for instance, correspond to the ternary string $\texttt{txt} = \star 00 \star 011 \star\star 1110 \star 1$, with the same interpretation of the attributes as before. Note that the values of some attributes are $\star$; we will say that these attributes are ***masked in*** $\texttt{txt}$. Observe that this particular piece of text *does not* state whether the bank robber was holding a gun (the value of x_1 in $\texttt{txt}$ is $\star$), but it does say that the bank robbery was *not* successful (the value of x_2 in $\texttt{txt}$ is 0). Thus, this piece of text is not accurately depicting the underlying reality according to $\texttt{evn}$, since the actual bank robbery was successful. We will say that x_2 is ***noisy in*** $\texttt{txt}$ ***w.r.t.*** $\texttt{evn}$.

Recall that a reader hopes to reconstruct the actual event, given as input a piece of text (partially and noisily) describing it. In terms of our example, the reader is given $\texttt{txt} = \star 00 \star 011 \star\star 1110 \star 1$, and attempts to construct some new ternary string that resembles $\texttt{evn} = 010001110111001$ more closely than $\texttt{txt}$. But, is this task even well-defined? The reader does not have access to the actual event. It is possible that the same piece of text describes many bank robberies that differ amongst them. A bank robbery where a gun was involved, and one where a gun was not are equally well described by the text "The bank robber fled the scene with the loot.". Which of the two possible realities is the reader supposed to reconstruct? Or, in other words, does this piece of text entail the statement that "a gun was used during the bank robbery"? The answer depends on what was observed by the text's author! This response becomes useful when the author is modelled. The author is modelled as a (stochastic) process that takes as input events (i.e., binary strings), and outputs text (i.e., ternary strings). An event exists first, and is then mapped into text through the workings of an author. When reading this text the reader attempts to recover the event. We

will call any such process that maps binary to ternary strings a ***sensor***; the author is the sensor through which a reader observes an event. The reader's task is now well-defined.

To understand why it is even conceivable to be able to recover information about some underlying event, by only reading some text about it, one needs to appreciate the fact that the reader also employs some prior knowledge presumably relevant to the event of interest. Most readers know, for instance, that bank robberies often involve the use of some form of an assault weapon, and a means of escaping the crime scene. Even if the news reporter does not mention such information, it is reasonable to assume that the reader will infer it given the context. The prior knowledge employed by a reader, then, can be modelled as a rule that tries to predict the value of some target attribute x_t given only `txt`. We take the approach that this rule $(\varphi) \equiv x_t$ is of the typical form. The rule's body φ is a propositional formula defined over the attributes $\mathcal{A}$ of a domain, and the rule's head is the attribute x_t to be predicted. In our example, the rule could be of the following form: the bank robber was holding a gun and an accomplice was waiting outside the bank if and only if the bank robbery was successful. Given such a rule, and given a piece of text offering the information that indeed the bank robber was holding a gun and an accomplice was waiting outside the bank, then the reader can infer that the bank robbery was successful, irrespectively of whether the text offers this information. If we are confident in the accuracy of the rule, then we should be willing to accept that its prediction is correctly reconstructing the particular event described by the given piece of text.

3 A Semantics for Textual Entailment

We denote by ϱ the knowledge base that an agent uses to recognize textual entailment, including both the actual knowledge, and the inference engine that applies this knowledge on a given piece of text `txt`. We write $\texttt{conc}(\varrho \,|\, \texttt{txt})$ to mean the ternary string that results when ϱ is applied on `txt`.

A first attempt to define textual entailment is to ask that for a sensing process `sense`, an event `evn`, and a piece of text `txt` drawn from $\texttt{sense}(\texttt{evn})$, it holds that $\texttt{conc}(\varrho \,|\, \texttt{txt})$ is equal to `evn`. Thus, the application of the knowledge base ϱ on the piece of text `txt` reconstructs the hidden event `evn`.

Clearly, for any knowledge base ϱ, an event `evn` can be adversarially chosen so as to make it impossible to reliably (i.e., not by chance) reconstruct `evn`, even if no attribute is noisy in `txt`, and only a single attribute is masked in `txt`. A minimal assumption is needed to circumvent this: that the events have some structure, in the sense that the values of the attributes in an event are somehow correlated with each other. Following the approach of standard learning models (e.g., [Valiant, 1984]), the possibly complex correlations that exist among the values of attributes in the underlying reality are captured by assuming that an event is drawn from some *arbitrary* probability distribution $\mathcal{D}$; we write $\texttt{evn} \leftarrow \mathcal{D}$. By the same token, some minimal structure should be assumed on the sensing process. We model the possibly complex workings of an author by assuming that once an event `evn` is chosen, a piece of text `txt` is drawn from some *arbitrary* probability distribution $\texttt{sense}(\texttt{evn})$; we write $\texttt{txt} \leftarrow \texttt{sense}(\texttt{evn})$.

We can now define two metrics that measure the extent to which some knowledge base ϱ succeeds in the textual entailment task. Soundness amounts to asking that if some attribute in a selected set $\mathcal{A}_t$ has a $\{0, 1\}$ value in $\texttt{conc}(\varrho \,|\, \texttt{txt})$, then this value should match the one specified by the underlying event `evn` that gave rise to `txt`. Completeness amounts to asking that the attributes in $\mathcal{A}_t$ should have a $\{0, 1\}$ value in $\texttt{conc}(\varrho \,|\, \texttt{txt})$. For flexibility, we ask that soundness and completeness hold except with some small probability, giving them a statistical flavor, without sacrificing objectivity.

Definition 1 (Soundness) *A knowledge base ϱ is $(1-\varepsilon)$-**sound for** a target attribute set $\mathcal{A}_t \subseteq \mathcal{A}$ **under** a probability distribution $\mathcal{D}$ **and** a sensing process* `sense` *if*

$$Pr\,[\exists x_t \in \mathcal{A}_t \textit{ that is noisy in } \texttt{conc}(\varrho \,|\, \texttt{txt}) \textit{ w.r.t. } \texttt{evn} \mid \texttt{evn} \leftarrow \mathcal{D}; \texttt{txt} \leftarrow \texttt{sense}(\texttt{evn})] \leq \varepsilon.$$

Definition 2 (Completeness) *A knowledge base ϱ is $(1-\omega)$-**complete for** a target attribute set $\mathcal{A}_t \subseteq \mathcal{A}$ **under** a probability distribution $\mathcal{D}$ **and** a sensing process* `sense` *if*

$$Pr\,[\exists x_t \in \mathcal{A}_t \textit{ that is masked in } \texttt{conc}(\varrho \,|\, \texttt{txt}) \mid \texttt{evn} \leftarrow \mathcal{D}; \texttt{txt} \leftarrow \texttt{sense}(\texttt{evn})] \leq \omega.$$

We emphasize that the soundness and completeness metrics differ from the precision and recall metrics. Both pairs of metrics are employed in cases where some universe of objects (in our case, these are events) are classified as having some property or not (in our case, this is the value of a single fixed attribute), and the goal is to accurately predict what the case is for each object. In the case of precision and recall, predictions are made on *every* object, corresponding to having a 1-complete knowledge base. By contrast, we allow for knowledge bases that do not make predictions on certain attributes, making them $(1-\omega)$-complete for $\omega > 0$. Another important difference is the asymmetry that precision and recall impose between *having* the property of interest or not. Precision is: out of the objects predicted to have the property, how many actually do? Recall is: out of the objects that have the property, how many were predicted as such? Soundness, on the other hand, measures how often the value of an attribute was not predicted wrongly (i.e., it was either predicted correctly or not predicted at all). There is no asymmetry in the treatment of the $\{0, 1\}$ values that an attribute may take.

Precision and recall are often employed for evaluating success in the textual entailment task. In such evaluations, a predictor is given a piece of text and a statement and is asked to determine whether the latter follows from the former. Our proposed view of textual entailment goes beyond treating it as the *classification* task of checking whether a statement follows from a piece of text. Instead, we view textual entailment more broadly as the *generation* task of drawing all possible inferences that can be reliably drawn from a given piece of text. The traditional classification view of textual entailment is, then, a special case: check if the statement is one of the drawn inferences. In our setting, soundness and completeness are more appropriate metrics than precision and recall.

4 Learning Background Knowledge

Manually constructed knowledge bases (e.g., [Lenat, 1995]), are not accompanied by guarantees on their appropriateness

for any particular textual entailment task. Learning has been used to construct knowledge bases for the textual entailment classification task, yet, most approaches proceed by processing a training set comprised of pairs of a piece of text and a statement, tagged — in line with the operational and subjective definition of textual entailment — by *humans* to indicate whether the latter is implied by the former. For autonomous machines to be built, and for the broader textual entailment *generation* task to be solved, learning techniques not relying on human supervision need to be employed. Access to text is still allowed, but text is not tagged with the inferences expected to be drawn. Since text is a partial depiction of some underlying reality, a learning framework that can deal with partial information in its learning examples is needed. It is also desirable that such a learning framework makes minimal assumptions on how information is missing in its learning examples, allowing it thus to be employed as broadly as possible. We review in this section one such learning framework, called *autodidactic* [Michael, 2007]. The reader is directed to the cited paper for details and discussion of related work.

The autodidactic learning framework extents the Probably Approximately Correct semantics [Valiant, 1984], and inherits from that the formal guarantees for learned knowledge. On the other hand, it can deal with partial learning examples, and does without the supervised nature of the original PAC model. For this section, we focus our attention on how a single rule for predicting a fixed target attribute x_t is to be learned.

Consider the domain of bank robberies, and let x_t stand for "the bank robbery was successful". What does it mean when we say that a rule $(\varphi) \equiv x_t$ is appropriate for inferring from text whether x_t is true or not? Following Definition 1, this could be formalized as asking that the rule does not wrongly predict the value of x_t in some underlying event $\texttt{evn}$, given access to a piece of text $\texttt{txt} \leftarrow \texttt{sense(evn)}$. Note that completeness is taken into account implicitly, since the rule makes a "don't know" prediction if and only if $\texttt{txt}$ does not offer sufficient information for φ to be uniquely determined. Denote by $\texttt{val}(\varphi \,|\, \texttt{txt})$ the prediction that the rule $(\varphi) \equiv x_t$ makes on the value of its head x_t given a piece of text $\texttt{txt}$.

Say that $(\varphi) \equiv x_t$ has an ***accuracy conflict with*** $\texttt{evn}$ ***given*** $\texttt{txt}$ if $\texttt{val}(\varphi \,|\, \texttt{txt}) \in \{0, 1\}$ and $\texttt{val}(\varphi \,|\, \texttt{txt}) \neq \texttt{evn}[t]$.

So, for the rule $(x_3 \vee (x_1 \wedge \overline{x_7})) \equiv x_4$, the event $\texttt{evn} = \texttt{1000110}$, and the text $\texttt{txt} = \texttt{10***10}$, it holds that $\texttt{val}(x_3 \vee (x_1 \wedge \overline{x_7}) \,|\, \texttt{txt}) = 1$, and $\texttt{evn}[4] = 0$; there is an accuracy conflict, since the predicted value of x_4 is incorrect.

Definition 3 *A rule $(\varphi) \equiv x_t$ is $(1-\varepsilon)$-**accurate under** a probability distribution $\mathcal{D}$ **and** a sensing process* $\texttt{sense}$ *if*

$$Pr\,[(\varphi) \equiv x_t \text{ has an accuracy conflict with } \texttt{evn} \text{ given } \texttt{txt} \mid \texttt{evn} \leftarrow \mathcal{D}; \texttt{txt} \leftarrow \texttt{sense(evn)}] \leq \varepsilon.$$

How is a rule evaluated when the underlying reality it is trying to infer and against which its accuracy is measured is unknown? Interestingly, it suffices to consider another metric.

Say that $(\varphi) \equiv x_t$ has a ***consistency conflict with*** $\texttt{txt}$ if $\texttt{val}(\varphi \,|\, \texttt{txt}), \texttt{txt}[t] \in \{0, 1\}$ and $\texttt{val}(\varphi \,|\, \texttt{txt}) \neq \texttt{txt}[t]$.

So, for the rule $(x_3 \vee (x_1 \wedge \overline{x_7})) \equiv x_4$, and the text $\texttt{txt} = \texttt{10***10}$, it holds that $\texttt{val}(x_3 \vee (x_1 \wedge \overline{x_7}) \,|\, \texttt{txt}) = 1$, and $\texttt{txt}[4] = *$; there is no consistency conflict independently of the value of x_4 in the underlying event that gave rise to $\texttt{txt}$.

Definition 4 *A rule $(\varphi) \equiv x_t$ is $(1-\varepsilon)$-**consistent under** a probability distribution $\mathcal{D}$ **and** a sensing process* $\texttt{sense}$ *if*

$$Pr\,[(\varphi) \equiv x_t \text{ has a consistency conflict with } \texttt{txt} \mid \texttt{evn} \leftarrow \mathcal{D}; \texttt{txt} \leftarrow \texttt{sense(evn)}] \leq \varepsilon.$$

The next result shows that highly consistent rules are also highly accurate *to the extent possible* [Michael, 2007].

Theorem 4.1 (The Relation of Consistency and Accuracy) *For every noiseless sensing process* $\texttt{sense}$*, and every class $\mathcal{F}$ of rules with head x_t, there exists $\eta \in [0, 1]$ s.t. if $\eta \neq 0$:*

(i) for every probability distribution $\mathcal{D}$, and rule $(\varphi) \equiv x_t$, the rule is $(1-\varepsilon)$-accurate if it is $(1-\eta \cdot \varepsilon)$-consistent;

(ii) there is a probability distribution $\mathcal{D}_0$, and a rule $(\varphi_0) \equiv x_t$ that is $(1-\varepsilon)$-accurate only if it is $(1-\eta \cdot \varepsilon)$-consistent.

Roughly, $1-\eta$ measures how adversarially the author of text (on which rules in $\mathcal{F}$ are applied) hides information.

An immediate implication for the textual entailment task is that in order to be able to infer whether "the bank robbery was successful", and be confident that this was indeed the case in the actual event, it suffices to construct a rule that when tested on pieces of text, it is almost never found to be in conflict *with the text*; it either makes no prediction due to lack of information to determine whether the rule's premises hold, or it makes a prediction but the text does not offer any information of whether this is true or false, or it makes a prediction that is corroborated by the text. Assuming that such a rule can be found, this rule is, by virtue of Theorem 4.1, reliable in the sense that it will almost never predict something that is in conflict *with the actual event* underlying the text. Thus, such rules, learned from text alone, can be later applied on new pieces of text and reliably, in a precisely defined sense, recover information *on the event observed by the author*. Sufficient learnability conditions are given next [Michael, 2007].

Theorem 4.2 (Autodidactic Learning from Text) *Assume: (i) in the underlying reality, as determined by some arbitrary probability distribution $\mathcal{D}$, the value of some target attribute x_t is determined by some monotone formula ψ over the rest of the attributes; (ii) events drawn from $\mathcal{D}$ are mapped into text through an arbitrary noiseless sensing process* $\texttt{sense}$*; (iii) the formula ψ belongs in a class of formulas that is learnable in the standard PAC model (from complete examples). Then: there exists an algorithm that given access to pieces of text drawn from* $\texttt{sense}(\mathcal{D})$*, runs in time polynomial in the relevant learning parameters, and returns a rule $(\varphi) \equiv x_t$ that, w.h.p., is $(1-\varepsilon)$-consistent under $\mathcal{D}$ and* $\texttt{sense}$*.*

From known PAC results [Blumer *et al.*, 1989], the class of linear threshold formulas is learnable in the sense above. This is a rather broad and useful class of rules, employed in practice for knowledge acquisition in many machine learning frameworks (see, e.g, [Roth and Yih, 2001; Valiant, 2006]).

5 How Should Learned Rules be Used?

Given rules that have been learned, how should they be used for recovering missing information implied by a piece of text? Conceivably, an inference engine could choose to apply all rules in parallel, that is, check whether the premises of each

rule are satisfied given the original text. Some other inference engine could choose to apply some of the rules, and extent the piece of text with the rules' conclusions. The remaining rules could be then applied on this extended piece of text. Thus, this second layer of rules would be able to take into account the conclusions of the rules in the first layer. Are the two approaches equally beneficial for the textual entailment task?

Valiant [2006] offers pragmatic considerations in favor of a layered application of rules in the context of automated acquisition and handling of unaxiomatized knowledge: the statistics of the data might not support the induction of rules within a single layer, and even if they do, the induction task might be computationally hard; and programmed rules might need to be integrated in the reasoning process (naturally accommodated by applying them in a layer prior to the learned rules).

Certain empirical evidence discussed by Dietterich [2000] in the context of aggregating multiple *learned* rules could also be viewed as supporting the same conclusion: a *statistical* reason relates to the scarcity of training data; a *computational* reason appeals to the hardness of searching the hypothesis space; and a *representational* reason accounts for the case that the hypothesis class is not expressive enough.

In the context of textual entailment, we are able to formally show that applying rules in multiple layers is *provably* beneficial. This result holds without appealing to any statistical, computational, or representational assumptions, and irrespectively of whether the rules are learned or programmed.

Say that ***reasoning collapses for*** a target attribute set $\mathcal{A}_t \subseteq \mathcal{A}$ ***under*** a probability distribution $\mathcal{D}$ ***and*** a sensing process `sense` if for every $(1-\varepsilon)$-sound and $(1-\omega)$-complete knowledge base, there exists a $(1-\varepsilon')$-sound and $(1-\omega')$-complete *single-layered* knowledge base s.t. $\varepsilon'+\omega' \leq \varepsilon+\omega$.

Thus, reasoning does not collapse if it is possible to find a knowledge base that *strictly* outperforms, in terms of soundness and completeness, *every* knowledge base whose rules are applied in parallel. We claim that the knowledge base that chains the following two rules in two layers has this property:

Layer 1 Rule: (*noon time*) $\equiv$ *lunch time*.

Layer 2 Rule: (*lunch time* and *Bob hungry*) $\equiv$ *Bob eats*.

The setting is one where an agent learned the two rules above by reading text. It now faces the new piece of text "It was not noon time yet, but Bob was hungry." and wishes to determine whether this text implies *Bob eats*. Applying the rule in the first layer yields that is is not *lunch time*. Applying, then, the rule in the second layer yields that *Bob eats* does not hold. Consider now an attempt to merge the two rules into a single rule. One such attempt, for instance, could be the following:

((*noon time* or *lunch time*) and *Bob hungry*) $\equiv$ *Bob eats*.

Note that this rule *makes no prediction*, since the text does not offer sufficient information to uniquely determine the rule's premises; *lunch time* is unknown. A similar phenomenon can be reproduced for any choice of a single rule[1], and even if the given text is "It was not lunch time yet, but Bob was hungry.".

This phenomenon manifests itself whenever more than one piece of information unilaterally determines what holds in some context. In our example, the context is that Bob is hungry, and the two pieces of information each of which unilaterally determines that *Bob eats* are *noon time* and *lunch time*.

Consider a rule $(\varphi) \equiv x_t$, and an attribute x_i that is masked in a piece of text `txt`. If changing the value of x_i to `0` or `1` causes the rule to make different $\{0,1\}$ predictions, then call x_i ***critical for*** $(\varphi) \equiv x_t$ ***w.r.t.*** `txt`. It is not hard to show:

Lemma 5.1 (Unique Critical Attribute) *At most one attribute is critical for any rule* $(\varphi) \equiv x_t$ *w.r.t. any text* `txt`.

Lemma 5.1 implies an inherent and unconditional limitation of individual rules. No rule encoding background knowledge about the underlying reality can have more than one source of information *unilaterally* determining its prediction. If such types of knowledge need to be encoded, then, this should be done by applying multiple rules in multiple layers.

A sensing process `sense` is ***k-critical inducing under*** a probability distribution $\mathcal{D}$ ***for*** a target attribute x_t, if in some text drawn from $\texttt{sense}(\mathcal{D})$ with non-zero probability, each of k attributes unilaterally determines[2] the value of x_t.

Theorem 5.2 (Domains with Non-Collapsible Reasoning) *For every* $k \geq 2$*, and every noiseless sensing process* `sense` *that is* k*-critical inducing under a probability distribution* $\mathcal{D}$ *for a target attribute* x_t*, reasoning does not collapse for* $\{x_t\}$ *under* $\mathcal{D}$ *and* `sense`.

Proof (sketch): Choose ϱ_1 that maximizes the performance among knowledge bases that use only a single layer. Extend it to ϱ_k by adding perfectly accurate rules to predict x_t from the values of the k attributes that determine the value of x_t. The multi-layered knowledge base ϱ_k can be shown to be perfectly sound and complete on some particular set $\mathcal{O}$ of texts. As a consequence of Lemma 5.1, any knowledge base that uses a single layer and makes $\{0,1\}$ predictions on all texts in $\mathcal{O}$, necessarily makes at least one wrong prediction. □

6 The Necessity of Learning in Layers

The formal result that reasoning does not collapse in the context we consider, can be seen as a prescription that it is beneficial to chain pieces of knowledge when engaged in the textual entailment task. Theorem 5.2, however, shows only that there exists some beneficial way to chain rules; it does not say how an appropriate order in which to apply the rules can be identified. It can be easily seen that not every ordering of rules is (equally) useful. In our earlier example for predicting whether *Bob eats*, chaining the two rules in the reverse order would result in the rule that predicts *Bob eats* not exploiting the conclusions of the rule that predicts *lunch time*.

Before attempting to determine how to order rules, we ask a more basic question. Can the rules be learned independently of each other, and then be ordered? That is, can the learning of rules be decoupled from the way they are ordered for drawing inferences? An easy argument shows that this is not the

[1] Note that rules of the form "((it is noon time and *the text does not state* that it is not lunch time) and Bob is hungry) if and only if Bob eats" are not permissible. Such rules do not encode background knowledge about the underlying reality. Instead, they try to encode the way that the text's author omits information when writing.

[2] In the sense discussed above, where realities that differ on the value of any one of these attributes, also differ on the value of x_t.

case. Indeed, assume that multiple rules are learned independently, and that they are subsequently chained in a knowledge base. Consider a rule $(\varphi) \equiv x_t$ in the last layer of the constructed knowledge base. Although during the learning phase $(\varphi) \equiv x_t$ faces as inputs original pieces of text (i.e., outputs of some fixed sensing process `sense`), during the reasoning phase the rule faces *different* pieces of text, namely those that result when the original pieces of text are extended with the inferences drawn from rules in earlier layers of the knowledge base. It is easy to construct adversarial scenarios where this change in the pieces of text given as inputs to $(\varphi) \equiv x_t$ causes the rule to suffer an arbitrary loss in accuracy with respect to what it achieves when given the original pieces of text.

This argument suggests that rules in higher reasoning layers should be learned by facing the same inputs that they will face during the reasoning phase. This can be achieved only if when learning $(\varphi) \equiv x_t$, all rules in the previous layers have already been learned, and are applied on pieces of text to draw inferences, which are then given as input (along with the original pieces of text) to $(\varphi) \equiv x_t$, which uses them as training instances. Thus, an iterative learning strategy that interleaves learning and reasoning is necessitated. Such a strategy is beneficial, in that the multi-layered knowledge bases it constructs improve upon the single-layered knowledge bases.

Theorem 6.1 (Iterative Learning) *By interleaving learning and reasoning, and by an appropriate choice of accuracies for rules at each layer, a learning algorithm $\mathcal{L}$ as those whose existence is guaranteed by Theorem 4.2 can be extended to one that produces multi-layered knowledge bases, does not sacrifice the soundness of knowledge bases returned by $\mathcal{L}$, and in certain cases improves upon their completeness.*

Proof (sketch): Before learning the rules in the j-th layer, ensure that the knowledge base ϱ_{j-1} that is comprised of the $j-1$ first layers is $(1-\varepsilon_{j-1})$-sound for a sufficiently small ε_{j-1}, so that, w.h.p., no unsound inference will be drawn during the training phase of the rules in the j-th layer. This requirement propagates recursively to lower layers, and is amplified each time in a manner that grows exponentially with j. To ensure efficiency, we restrict j to a constant. □

Our study [Michael, 2008] has shown that the guarantees offered by the aforementioned iterative learning strategy cannot, in general, be surpassed by other natural approaches.

7 Conclusions

An objective metric for measuring success in the textual entailment task was proposed, and was shown to naturally accommodate a generalization of textual entailment from a classification task to a generation task. It was argued that machines that completely autonomously engage in the textual entailment generation task can be built by exploiting existing work in Machine Learning. The case of using multiple rules was examined, and it was shown that it is beneficial for such rules to be chained in multiple layers, and that rules should not be learned independently, but in an iterative manner.

This work provides a formal basis for actual systems to be built that engage in textual entailment. Their ultimate success is determined in part by an appropriate selection of learning features. Traditional NLP tasks can aid in extracting information encoded in text, in terms of propositional (or relational) statements. Collections of such statements will correspond to the ternary strings `txt` used in our framework. This is not to suggest that textual entailment cannot, itself, aid in enhancing the performance of systems for traditional NLP tasks. Indeed, an iterative approach, along the lines of that presented in Section 6, seems to provide a fruitful and mutually beneficial interaction of traditional NLP tasks and textual entailment.

Some of the presented ideas have been employed in experimental work that examined the feasibility of massive knowledge infusion [Michael and Valiant, 2008]. Although the task considered in that work is broader than textual entailment, the conclusions of that work suggest that our proposed approach to textual entailment is applicable in real-world settings.

Acknowledgements

This work has benefited from discussions with Leslie Valiant.

References

[Blumer *et al.*, 1989] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *JACM*, 36(4):929–965, 1989.

[Dagan *et al.*, 2005] I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognizing Textual Entailment Challenge. In *RTE'05*, 2005.

[Dietterich, 2000] T. G. Dietterich. Ensemble Methods in Machine Learning. In *MCL'00*, 2000.

[Etzioni *et al.*, 2006] O. Etzioni, M. Banko, and M. J. Cafarella. Machine Reading. In *AAAI'06*, 2006.

[Hickl *et al.*, 2006] A. Hickl, J. Williams, J. Bensley, K. Roberts, B. Rink, and Y. Shi. Recognizing Textual Entailment with LCCs Groundhog System. In *RTE'06*, 2006.

[Lenat, 1995] D. B. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. *CACM*, 38(11):33–38, 1995.

[Michael and Valiant, 2008] L. Michael and L. G. Valiant. A First Experimental Demostration of Massive Knowledge Infusion. In *KR'08*, 2008.

[Michael, 2007] L. Michael. Learning from Partial Observations. In *IJCAI'07*, 2007.

[Michael, 2008] L. Michael. *Autodidactic Learning and Reasoning*. PhD thesis, School of Engineering and Applied Sciences, Harvard University, U.S.A., 2008.

[Mitchell, 2005] T. M. Mitchell. Reading the Web: A Breakthrough Goal for AI. *AI Magazine*, Fall 2005.

[Roth and Yih, 2001] D. Roth and W. Yih. Relational Learning via Propositional Algorithms: An Information Extraction Case Study. In *IJCAI'01*, 2001.

[Valiant, 1984] L. G. Valiant. A Theory of the Learnable. *CACM*, 27(11):1134–1142, 1984.

[Valiant, 2000] L. G. Valiant. Robust Logics. *Artificial Intelligence*, 117(2):231–253, 2000.

[Valiant, 2006] L. G. Valiant. Knowledge Infusion. In *AAAI'06*, 2006.