Monte Carlo Semantics: MCPIET at RTE4

Richard Bergmair
recipient of a DOC-fellowship of the Austrian Academy of Sciences
at the University of Cambridge Computer Laboratory;
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK;
rbergmair@acm.org

Abstract

The Monte Carlo Pseudo Inference Engine for Text (MCPIET) addresses the RTE problem within a new theoretic framework for robust inference and logical pattern processing based on integrated deep and shallow semantics.

In this report we outline, in some detail, this new theoretic framework, and we will use it to shed some light on the informativity and robustness characteristics for the extreme cases of deep and shallow processing. Unsurprisingly, it will turn out that there is a tradeoff between informativity and robustness.

We will be able to characterize an important new notion of a degree of validity, and provide some evidence to suggest that this concept plays a crucial role in the robustness of shallow inference. At the same time our framework still supports informationally rich semantic representations and background theories, which play the central role in the informativity of deep inference.

Within our new theory we can then pose, from a completely new perspective, the problem of deep/shallow integration, and also propose a solution to it, which we will call *Monte Carlo Semantics*.

1 Overview

Informativity is the ability of a system to take into account all available relevant information.

Robustness is the ability of a system to proceed on reasonable assumptions, where relevant information is missing.

Many current techniques for RTE can roughly be situated anywhere along a spectrum between deep and shallow techniques, and suffer from a tradeoff between informativity and robustness.

Deep techniques, like the NUTCRACKER[1] RTE-system (Bos and Markert, 2005; Bos and Markert, 2006), are informative but not very robust. Shallow techniques, like a bag-of-word overlap measure, are robust but not very informative.

[1] NUTCRACKER attempts to translate text into logic, and use theorem provers to reason about text within a logic and a given theory of background knowledge.

The great majority of systems are perhaps to be regarded as intermediate-level techniques, situated somewhere between deep and shallow. For example the LOGINF system (MacCartney and Manning, 2007; Chambers et al., 2007; MacCartney and Manning, 2008), works directly off the statistical Stanford parser. Such intermediate-level systems provide intermediate levels of both informativity and robustness.

Deep/Shallow integration tries to escape the informativity/robustness tradeoff altogether, by combining into a single technique the informativity of deep methods and the robustness of shallow methods.

Prior work by (Bos and Markert, 2005; Bos and Markert, 2006) suggests that this cannot be achieved by simply running deep and shallow systems independently and expecting a machine learner to determine which of the two is right in any given situation.

Rather, we believe, that this integration can best be achieved, by first formulating a unified theory of textual inference. Properties of informativity and robustness must be understood within such a theory, and current deep and shallow techniques should turn out to be expressible within such a unified theoretic framework. Deep/shallow integration might then be achieved by formulating a single technique within the unified theory in all its generality.

In the rest of this paper, we will outline the core of such a unified theoretic framework, and we will report on our first steps towards developing such an integrated single technique.

Our point of departure for this theoretic framework is traditional logic. Yet, within our essentially logical approach, we advocate an important paradigm shift. Where logic is traditionally concerned with *whether or not* a given candidate entailment "$\varphi \rightarrow \psi$" is valid, MCPIET uses a new Monte Carlo technique to *estimate*, in a probabilistic sense, the *degree* of validity for such a formula. It is this more flexible and relative notion of a degree of validity which lends important robustness characteristics to the logical approach. This is the core of our contribution, and the rest of this paper will give an outline of how it is defined, and why it is useful to RTE.

1.1 Deep, Shallow, and their Limitations

To illustrate the problem more concretely, let us consider some of the examples in figure 1. For now, let $\top$ stand for *valid*, $\bot$

(1) predicate/argument structures

(a) $\quad \top \;>\; \dfrac{\text{The cat chased the dog.}}{\rightarrow \quad \text{The dog chased the cat.}}$

(2) monotonicity properties, upwards entailing

(a) $\quad \dfrac{\text{Some (grey } X\text{) are } Y}{\rightarrow \quad \text{Some } X \text{ are } Y} \;\geq\; \top$

(b) $\quad \top \;>\; \dfrac{\text{Some } X \text{ are } Y}{\rightarrow \quad \text{Some (grey } X\text{) are } Y}$

(c) $\quad \dfrac{\text{Some } X \text{ are } Y}{\rightarrow \quad \text{Some (grey } X\text{) are } Y} \;>\; \dfrac{\text{Some } X \text{ are } Y}{\rightarrow \quad \text{Some (clean (grey } X\text{)) are } Y}$

(3) monotonicity properties, downwards entailing

(a) $\quad \dfrac{\text{All } X \text{ are } Y}{\rightarrow \quad \text{All (grey } X\text{) are } Y} \;\geq\; \top$

(b) $\quad \top \;>\; \dfrac{\text{All (grey } X\text{) are } Y}{\rightarrow \quad \text{All } X \text{ are } Y}$

(c) $\quad \dfrac{\text{All (grey } X\text{) are } Y}{\rightarrow \quad \text{All } X \text{ are } Y} \;>\; \dfrac{\text{All (clean (grey } X\text{)) are } Y}{\rightarrow \quad \text{All } X \text{ are } Y}$

(4) quantifiers and negations

(a) $\quad \top \;>\; \dfrac{\text{Some } X \text{ are } Y}{\rightarrow \quad \text{All } X \text{ are } Y}$

(b) $\quad \bot \;\geq\; \dfrac{\text{Some } X \text{ are } Y}{\rightarrow \quad \text{No } X \text{ are } Y}$

(c) $\quad \bot \;\geq\; \dfrac{\text{All } X \text{ are } Y}{\rightarrow \quad \text{Some } X \text{ are not } Y}$

(6) sentential connectives on clauses)

(a) $\quad \dfrac{\text{S is a man and every man is mortal}}{\rightarrow \quad \text{S is mortal}} \;\geq\; \top$

(7) gradual standards of proof (tautology $>$ contingency $>$ contradiction)

(a) $\quad \dfrac{\text{Socrates is a man}}{\rightarrow \quad \text{Socrates is a man}} \;>\; \dfrac{\text{Socrates is a man}}{\rightarrow \quad \text{Socrates is mortal}}$

(b) $\quad \dfrac{\text{Socrates is a man}}{\rightarrow \quad \text{Socrates is mortal}} \;>\; \dfrac{\text{Socrates is a man}}{\rightarrow \quad \text{Socrates is not a man}}$

Figure 1: example inferences

for *unsatisfiable*. So $\chi \geq \top$ means that the candidate entailment χ is valid (ENTAILMENT), and $\bot \geq \chi$ means that χ is unsatisfiable (CONTRADICTION). More generally, $\varphi > \psi$ means that φ is more valid than ψ. We will have much more to say about what exactly that means.

In example (1.a.), information about the predicate-argument structures of the texts is available. A shallow technique like bag-of-words overlap might miss out on this information and incorrectly decide that the entailment is true. Intermediate-level or deep approaches would likely get this example right. Hence the shallow method, in this example, fails on informativity, where deep methods would succeed.

Next, consider examples (2.c.) and (3.c.), substituting "elephants" for X. All elephants are grey, but not all elephants are clean. The antecedent fails to mention this, and it might well be the case that the system does not have access to this kind of information, in the face of an incomplete theory of real-world knowledge and common sense. Hence, we are missing relevant information. A deep technique would find that neither is the left-hand side candidate entailment provable nor is the right-hand side entailment. It would then decide that they are equally to be considered non-valid. But this is incorrect. Given that all elephants are grey, if some elephants are intelligent, then some grey elephants are intelligent, yet it is not true that some clean grey elephants are intelligent. Intermediate-level approaches or shallow approaches would easily get this distinction right. Each adjective illegally inserted in this way would contribute an additional penalty score to the entailment. Even if we don't know anything about elephants and under what conditions they can be considered clean or grey, it is still reasonable to assume that the insertion of *only one* adjective into the wrong slot of a quantifier is preferable to the illegal insertion of *two* such adjectives. – We could proceed on reasonable assumptions where we are missing relevant information, yet a traditional theorem prover will not do that. Hence, the deep method, in this example, fails on robustness, where shallow methods would succeed.

Intermediate-level systems comparing the symbolic structures of dependency parses seem to be both robust and informative, for both of these examples. Yet they would likely fail on other examples such as (2.a.), (3.a.), and (4), requiring a proper treatment of quantification. Examples such as (6) go even further, and require a logic to build on these quantified structures.

The LOGINF system (MacCartney and Manning, 2007; Chambers et al., 2007; MacCartney and Manning, 2008), for example, is an attempt to engineer theoretical properties such as the ones exemplified in figure 1 into an intermediate-level system. While LOGINF does have its limitations, we should, perhaps, point out that our work provides a richer theory for this kind of inference, while theirs currently works better in practice. In fact, we believe that our theory and their system complement each other quite well.

Concerning MCPIET, we can now say, quite simply, that our goal is to get all of the theoretical properties right that are exemplified in figure 1.

1.2 Degree of Validity

We have already mentioned that it is our new notion of a degree of validity, which lends robustness characteristics like (2.c.) and (3.c.) to our theory. To see what exactly is meant by a degree of validity here, and how it relates to the problem of missing information, let us consider example (7) about the following three propositions:

$$\begin{aligned} \varphi: &\quad \text{“Socrates is a man”,} \\ \neg\varphi: &\quad \text{“Socrates is not a man”,} \\ \psi: &\quad \text{“Socrates is mortal”.} \end{aligned}$$

We are given some information, represented logically in a theory T. It is simply a set of formulae which we can assume to be valid a priori. If, on the basis of such an assumption, we must conclude that some formula χ is also valid, we say that χ is valid within theory T, written $T \models \chi$.

The well-known deduction theorem now defines when exactly a candidate entailment of the form "$\varphi \rightarrow \psi$" is valid. It states that $T \models \varphi \rightarrow \psi$ whenever $T \cup \{\varphi\} \models \psi$. In words: If we assume that the antecedent φ is valid, on top of all the formulae already in T, and we must conclude that ψ is also valid, then we also know that the candidate entailment "$\varphi \rightarrow \psi$" is valid in T.

When evaluating a given candidate entailment, there are traditionally four cases to distinguish:

(i) $T \cup \{\varphi\} \models \psi$ and $T \cup \{\varphi\} \not\models \neg\psi$;

(ii) $T \cup \{\varphi\} \not\models \psi$ and $T \cup \{\varphi\} \models \neg\psi$;

(iii) $T \cup \{\varphi\} \models \psi$ and $T \cup \{\varphi\} \models \neg\psi$;

(iv) $T \cup \{\varphi\} \not\models \psi$ and $T \cup \{\varphi\} \not\models \neg\psi$.

Assuming the empty theory $T = \emptyset$, the first candidate entailment in example (7), which is "$\varphi \rightarrow \varphi$", falls under case (i). The third candidate entailment, which is "$\varphi \rightarrow \neg\varphi$", falls under case (ii).

It is also quite common to require that a given theory $T \cup \{\varphi\}$ be *consistent*, i.e. that case (iii) does not occur.

But what about the second candidate entailment? Here we have "$\varphi \rightarrow \psi$", which would have to fall under case (iv). Here we are dealing with an incomplete theory.

In order to make the theory $T \cup \{\varphi\}$ *complete*, we could, for example, add

$$\chi: \quad \text{“Every man is mortal”}$$

to T. We would then have $\{\chi, \varphi\} \models \psi$ and $\{\chi, \varphi\} \not\models \neg\psi$, so the candidate entailment would fall under case (i). Or, we could have added χ', "No man is mortal", to make it fall under case (ii).

This is the major problem with NUTCRACKER, the major weakness with applying traditional theorem provers for classical logic to NLP tasks such as RTE: Theories are not, in practice, complete in that sense. They have to contain real-world knowledge and common sense.

Some knowledge of this kind is available. For example meaning postulates of the form "$\forall x : \text{cat}(x) \rightarrow \text{animal}(x)$" could be derived from the WordNet noun hyponymy hierarchy, or "$\forall x, y, z : \text{buy-from}(x, y, z) \equiv \text{sell-to}(z, y, x)$" could be derived from a role-labelled verb lexicon. Knowledge of more general kind might be automatically acquired from text. Given very careful knowledge engineering, one might even be able to ensure that the resulting theories are consistent. But assuming them to be complete in the above sense would be unrealistic at present.

In our opinion, such a completeness assumption, which would hold that case (iv) does not occur, is not only wrong, but, quite to the contrary, we would expect case (iv) to be indeed the most frequent one, with cases (i) and (ii) occuring only as limit cases of theoretical interest.

Thus, instead of talking about the bivalent dichotomy between validity and non-validity, we talk about what we call **degree of validity**. We write $T \models_t \chi$ iff χ is valid in T to a degree t from the rational-valued unit interval $[0, 1]$.

For example, χ could be valid in T, to a degree of 0.7, written $T \models_{0.7} \chi$.

We now distinguish the following cases:

(i) $T \cup \{\varphi\} \models_{1.0} \psi$ and $T \cup \{\varphi\} \models_{0.0} \neg\psi$;

(ii) $T \cup \{\varphi\} \models_{0.0} \psi$ and $T \cup \{\varphi\} \models_{1.0} \neg\psi$;

(iii) $T \cup \{\varphi\} \models_t \psi$ and $T \cup \{\varphi\} \models_{t'} \neg\psi$, for $0 < t, t' < 1.0$.

Here, everything is arranged along a continuum of degrees of validity. One can easily see, that this distinction is more fine-grained than the traditional dichotomy. Case (i) is simply the case of traditional validity, case (ii) is the case of traditional unsatisfiability, but in addition we have a new case (iii).

In this case (iii), we can now *compare* two given candidate entailments for their degree of validity, let us call them candidate 1, $T \cup \{\varphi_1\} \models_{t_1} \psi_1$, and candidate 2, $T \cup \{\varphi_2\} \models_{t_2} \psi_2$. It now may well be the case that we are missing knowledge, so that neither of them is strictly provable, in a proof-theoretic sense, that neither of them is a tautology, that neither of them is traditionally valid. But we can still determine, on the basis of the information we do have in T, which of them we would rather prove than the other, which of them is closer to being a tautology, which of them is valid to a higher degree. If $t_1 > t_2$, we prefer candidate 1, if $t_2 > t_1$, we prefer candidate 2.

Remarks

In section 3, we will explain, in greater detail, how we arrive at an actual number for the degree of validity of a given candidate entailment. For now, it is only important to note that, once we can compare candidate entailments on the basis of a degree of validity, we can, for example, sort the 800 candidate entailments in a given RTE dataset. Knowing that we expect 50% of them to be valid, we can quite naturally determine a cutoff to characterize when exactly a candidate entailment is good enough to be considered valid.

We can still distinguish the case of strict logical validity, so in section 3.1, we can set up traditional theorem proving as a special case of inference under our theory, showing that we can *potentially* gain the same level of informativity. Yet, we will generally find the notion of strict logical validity too restrictive a criterion to be useful in practice.

The more lenient criterion of using degrees of validity imposes a more useful structure on the truth classes in the case of a theory which would traditionally be considered incomplete, and thereby helps us to deal robustly with missing information. This allows us, in section 3.2, to set up bag-of-words overlap comparisons as another special case of textual inference, showing that we can *potentially* gain the same level of robustness.

Finally, in section 4, we will turn to the general case, and suggest a new technique to compute degrees of validity in this general case. We are then in a position to explain, what exactly was meant by the hedge *potentially* in the above two paragraphs in sections 4 and 5.

2 Logical Preliminaries

From external modules dealing with syntax and semantic composition, we expect a translation of the pieces of text T and H into formulae φ' and ψ' of a first-order predicate language. We then translate these into propositional logic, assuming that quantifications range over a finite domain of two individuals. For example "$\forall x : \text{P}(x) \rightarrow \text{Q}(x)$" would translate to "$(\text{p}_1 \rightarrow \text{q}_1) \wedge (\text{p}_2 \rightarrow \text{q}_2)$", and "$\exists x : \text{P}(x) \wedge \text{Q}(x)$" would translate to "$(\text{p}_1 \wedge \text{q}_1) \vee (\text{p}_2 \wedge \text{q}_2)$".

This leaves the problem of determining the degree of validity for the candidate entailment "$\varphi \rightarrow \psi$", now expressed as a formula in propositional logic. We approach this problem model-theoretically, i.e. by a process of considering its truth values under different conditions.

Definition 1. (truth values)

$$\mathbb{V}_2 \stackrel{\text{def}}{=} \{0, 1\}$$
$$\mathbb{V}_{\aleph_0} \stackrel{\text{def}}{=} \{v \mid v \in \mathbb{Q} \wedge 0 \leq v \leq 1\}.$$

So $\mathbb{V}_2$ is the set of two truth values used in classical logic. However, we will also need to make use of $\aleph_0$-valued Łukasiewicz logic[1] (Łukasiewicz and Tarski, 1930), which will be defined in greater detail later. For now, simply note that where classical logic assumes only the two truth values 0 and 1, a multi-valued logic could permit truth values like 0.7.

Definition 2. (propositional signature) We call Λ a *propositional signature*, iff Λ is a finite sequence of propositional symbols $\Lambda = \langle \text{p}_1, \text{p}_2, \ldots, \text{p}_N \rangle$ for some N.

Definition 3. (basic propositional syntax) The following recursive rules define by structured induction the notion of a *basic propositional formula over* $\mathbb{V}$ *and* Λ. For all v,p,φ, and ψ:

- if $v \in \mathbb{V}$, the *value constant* "$\overline{v}$" is a *formula*;
- if $p \in \Lambda$, the *proposition* "p" is a *formula*;
- if φ and ψ are *formulae*, then so is the *implication* "$(\varphi \to \psi)$";
- nothing else is a *formula*.

Definition 4. (extended propositional syntax) The following recursive rules define by structured induction the notion of an *extended propositional formula over* $\mathbb{V}$ *and* Λ. For all φ and ψ:

- If φ is a basic propositional formula over $\mathbb{V}$ and Λ, it is also an extended *formula*;
- If φ is a *formula*, then so is the *negation* "$\neg\varphi$";
- If φ and ψ are *formulae*, then so are the
  - *strong conjunction* "$(\varphi \,\&\, \psi)$",
  - *strong disjunction* "$(\varphi \veebar \psi)$",
  - *weak conjunction* "$(\varphi \wedge \psi)$",
  - *weak disjunction* "$(\varphi \vee \psi)$",
  - *equivalence* "$(\varphi \equiv \psi)$", and
  - *antivalence* "$(\varphi \not\equiv \psi)$";
- Nothing else is a *formula*.

Definition 5. [2] For any extended formula χ, we call χ' its *corresponding* basic formula, iff χ' results from χ by structured induction on the following transformation rules. For any formulae φ, ψ:

$$\begin{aligned}
\text{“}\neg\varphi\text{”} &\rightarrowtail \text{“}\varphi \to \overline{0}\text{”},\\
\text{“}\varphi \,\&\, \psi\text{”} &\rightarrowtail \text{“}\neg(\varphi \to \neg\psi)\text{”},\\
\text{“}\varphi \veebar \psi\text{”} &\rightarrowtail \text{“}\neg\varphi \to \psi\text{”},\\
\text{“}\varphi \wedge \psi\text{”} &\rightarrowtail \text{“}\varphi \,\&\, (\varphi \to \psi)\text{”},\\
\text{“}\varphi \vee \psi\text{”} &\rightarrowtail \text{“}(\varphi \to \psi) \to \psi\text{”},\\
\text{“}\varphi \equiv \psi\text{”} &\rightarrowtail \text{“}(\varphi \to \psi) \,\&\, (\psi \to \varphi)\text{”},\\
\text{“}\varphi \not\equiv \psi\text{”} &\rightarrowtail \text{“}\neg(\varphi \equiv \psi)\text{”}.
\end{aligned}$$

This defines the language of propositional logic. Its atomic symbols include a set of N propositional symbols $\Lambda = \langle \mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_N \rangle$. These have truth values that are in some sense variable. Furthermore, an atomic symbol could be a value constant like $\overline{0}$, $\overline{1}$, or $\overline{0.7}$, which would always have the truth values 0.0, 1.0, and 0.7, respectively.

Formulae are built out of such atomic symbols by combining them using operators like "$\to$", "$\wedge$", "$\neg$", etc. We take the implication operator "$\to$" and the value constant $\overline{0}$ as basic, in some sense, and define all other operators using only implications and the constant $\overline{0}$. Thus we will, from here on, always assume w.l.o.g. that formulae are in this basic syntax.

Now everything that remains to be done is to assign truth values to atomic symbols and a truth function to the operator "$\to$", and we have defined the truth value of any formula of propositional logic.

Definition 6. (valuation) We call w an N-dimensional *valuation* iff w is a vector

$$\vec{w} = [\, w_1 \; w_2 \; \ldots \; w_N \,]^{\mathrm{T}}.$$

We say w is bivalent, iff each $\mathrm{w}_i \in \mathbb{V}_2$, and that w is $\aleph_0$-valued otherwise, provided each $\mathrm{w}_i \in \mathbb{V}_{\aleph_0}$. We denote the set of all bivalent, N-dimensional valuations by $\mathcal{W}_{2,N}$ and the set of all $\aleph_0$-valued, N-dimensional valuations by $\mathcal{W}_{\aleph_0,N}$.

Definition 7. (classical propositional semantic) Let $\Lambda = \langle \mathsf{p}_1, \mathsf{p}_2, \ldots, \mathsf{p}_N \rangle$ be a propositional signature, and let w be a bivalent, N-dimensional valuation. Now, for any formula χ, the *truth value of* χ *over* w *and* Λ, denoted $\|\chi\|^{\Lambda}_{\mathrm{w}}$, is defined by structured induction as follows. For any formulae φ, ψ:

$$\begin{aligned}
\|\,\overline{0}\,\|^{\Lambda}_{\mathrm{w}} &\stackrel{\text{def}}{=} 0; \;\; \|\,\overline{1}\,\|^{\Lambda}_{\mathrm{w}} \stackrel{\text{def}}{=} 1;\\
\|\,\mathrm{p}_i\,\|^{\Lambda}_{\mathrm{w}} &\stackrel{\text{def}}{=} \mathrm{w}_i, \text{ for each } i;\\
\|\varphi \to \psi\|^{\Lambda}_{\mathrm{w}} &\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \|\varphi\|^{\Lambda}_{\mathrm{w}} = 1 \text{ and } \|\psi\|^{\Lambda}_{\mathrm{w}} = 1,\\ 0 & \text{if } \|\varphi\|^{\Lambda}_{\mathrm{w}} = 1 \text{ and } \|\psi\|^{\Lambda}_{\mathrm{w}} = 0,\\ 1 & \text{if } \|\varphi\|^{\Lambda}_{\mathrm{w}} = 0 \text{ and } \|\psi\|^{\Lambda}_{\mathrm{w}} = 1,\\ 1 & \text{if } \|\varphi\|^{\Lambda}_{\mathrm{w}} = 0 \text{ and } \|\psi\|^{\Lambda}_{\mathrm{w}} = 0. \end{cases}
\end{aligned}$$

Definition 8. (Łukasiewicz propositional semantic[1]) Let $\Lambda = \langle \mathsf{p}_1, \mathsf{p}_2, \ldots, \mathsf{p}_N \rangle$ be a propositional signature, and let w be a valuation. Now, for any χ, the *truth value of* χ *over* w *and* Λ, denoted $\|\chi\|^{\Lambda}_{\mathrm{w}}$, is defined by structured induction as follows. For any formulae φ, ψ:

$$\begin{aligned}
\|\,\overline{v}\,\|^{\Lambda}_{\mathrm{w}} &\stackrel{\text{def}}{=} v, \text{ for any } v \in \mathbb{V}_M;\\
\|\,\mathrm{p}_i\,\|^{\Lambda}_{\mathrm{w}} &\stackrel{\text{def}}{=} \mathrm{w}_i, \text{ for each } i;\\
\|\varphi \to \psi\|^{\Lambda}_{\mathrm{w}} &\stackrel{\text{def}}{=} \min(1, 1 - \|\varphi\|^{\Lambda}_{\mathrm{w}} + \|\psi\|^{\Lambda}_{\mathrm{w}}).
\end{aligned}$$

We have already defined that truth value constants always have particular truth values, so we would have $\|\overline{0.7}\| = 0.7$ regardless of the choice of a signature or valuation.

However, propositional symbols need to have their truth values assigned using a valuation. We might have $\Lambda = \langle \mathrm{p}_1, \mathrm{p}_2 \rangle$. Then a valuation might be $\vec{\mathrm{w}} = [\, 0.7 \; 0.3 \,]^{\mathrm{T}}$. This would assign to the propositional symbols the truth values $\|\mathrm{p}_1\|^{\Lambda}_{\mathrm{w}} = 0.7$ and $\|\mathrm{p}_2\|^{\Lambda}_{\mathrm{w}} = 0.3$.

If we use the implication operator to form a compound formula, then the truth value of that is determined from the truth values of the subformulae. For example $\|\mathrm{p}_1 \to \mathrm{p}_2\|^{\Lambda}_{\mathrm{w}} = \min(1, 1 - 0.7 + 0.3) = 0.6$.

Corollary 1. *Classical semantic (definition 7) is a special case of Łukasiewicz semantic (definition 8).*

This is straightforward, by substituting truth values 0 and 1 into the formula for the implication in definition 8. The truth values assigned to the implication by this formula will always coincide with the truth values assigned by the classical truth table.

Łukasiewicz logic is also complete w.r.t. modus ponens and the following axioms:

$$\begin{aligned} \varphi &\rightarrow (\psi \rightarrow \varphi); \\ (\varphi \rightarrow \psi) &\rightarrow \big((\psi \rightarrow \chi) \rightarrow (\varphi \rightarrow \chi) \big); \\ (\neg\varphi \rightarrow \neg\psi) &\rightarrow (\psi \rightarrow \varphi); \\ (\psi \vee \varphi) &\rightarrow (\varphi \vee \psi); \end{aligned}$$

Completeness proofs for this can be found throughout the relevant literature on multi-valued logic, (Rose and Rosser, 1958) being the earliest published proof of this kind[3].

From this proof-theoretic perspective, it is easy to see that, since all of these axioms are theorems of classical logic and since modus ponens is an inference rule in classical logic, Łukasiewicz logic will never prove a theorem not proved by classical logic. So, both from a proof-theoretic and model-theoretic perspective, it is clear that $\aleph_0$-valued Łukasiewicz logic is a generalization of bivalent logic.

Summary

It is perhaps important to stress, that we have had little to say in this section which is substantially new. We have simply introduced, for the convenience of the reader and for clarity of notation and terminology, some very basic concepts of logic. First, we mentioned a series of reductions of more complex logical notations to simpler ones. We convert text into predicate logic, predicate logic into propositional logic, and propositional logic into a basic syntax involving only implications and the value constant $\overline{0}$.

In order to define the truth value of any formula, it is then sufficient to provide a truth function for this implication operator. Traditional logic uses the two truth values $\{0, 1\}$ and the well-known truth table for implication. We also introduced Łukasiewicz logic as a generalization of this, which uses infinitely many truth values in the rational unit interval $[0, 1]$ and defines the implication operator as follows: $\|\varphi \rightarrow \psi\| = \min(1, 1 - \|\varphi\| + \|\psi\|)$.

3 Robust Informative Semantics

In section 1, we have already mentioned that our notion of a *degree of validity* is perhaps the core of our contribution. In section 1.2 we outlined some initial intuitions on this notion, and its function in defining a logic for robust natural language semantics. Now, with all the necessary logical preliminaries in place from section 2, we can move on to give a definition.

Definition 9. (degree of validity) Let $\Lambda = \langle \mathsf{p}_1, \mathsf{p}_2, \ldots, \mathsf{p}_N \rangle$ be a propositional signature, and let W be a set of N-dimensional valuations. The *degree of validity of a formula* χ *over* W *and* Λ, denoted $[\![\chi]\!]^{\Lambda}_{\mathrm{W}}$ is defined as follows:

$$[\![\chi]\!]^{\Lambda}_{\mathrm{W}} = \frac{1}{|\mathrm{W}|} \sum_{w \in \mathrm{W}} \|\chi\|^{\Lambda}_{w}.$$

In a first step, let us restrict attention to the case of considering all bivalent valuations, i.e. $\mathrm{W} = \mathcal{W}_{|\Lambda|,2}$. This definition can then easily be understood both logically and probabilistically.

For its logical interpretation, first recall the definitions of the classical notions of validity and satisfiability within such a model-theoretic framework. Here, χ is considered classically valid, iff the truth value $\|\chi\|_w$ equals 1 in *all* valuations w. We could also say, χ is classically valid iff the *minimum* truth value $\|\chi\|_w$ across all w is 1. Similarly, χ is considered classically satisfiable, iff $\|\chi\|_w$ equals 1 in *some* valuation w, i.e. iff the *maximum* truth value $\|\chi\|_w$ across all w is 1.

We have pointed out, before, that this traditional notion of validity is in practice too strong, and this notion of satisfiability too weak. A given candidate entailment will, in practice, usually turn out to be satisfiable, yet not valid. This is why we use a statistic between the minimum and maximum. We use an arithmetic mean.

Given a value for $[\![\chi]\!]$, we then know that χ is classically valid iff $[\![\chi]\!] = 1.0$, and that χ is classically satisfiable iff $[\![\chi]\!] > 0.0$. But we now also have a continuum of degrees of validity between the two extreme cases 1.0 and 0.0.

For its probabilistic interpretation, one can think of $\|\chi\|$ as a random variable indicating the truth value expected for $\|\chi\|_w$ when a valuation w is chosen from at random. The value of $[\![\chi]\!]$ is then quite simply the probability that the truth value of χ, for such a valuation w chosen at random, is 1, assuming for this choice a uniform distribution.

From the point of view of traditional objectivist probability, the question arises: Why should this distribution be uniform, rather than anything else? In response to this question, one might imagine an assumption of maximum entropy, i.e. maximum uncertainty, regarding this choice of a valuation.

From the point of view of subjectivist probability (De Finetti, 1974), which suits our theory much better, this question does not arise. The question, then, is not "Why assume a uniform distribution?", but rather "Why not?" – in the absence of any information contradicting such an assumption.

Another very interesting property of De Finetti's theory of probability is that it readily deals with the generalization where we move from the bivalent case $\mathrm{W} = \mathcal{W}_{N,2}$ to the more general case $\mathrm{W} = \mathcal{W}_{N,\aleph_0}$.

This does not have an interpretation in a frequentist view of probability, but the generalization is perfectly valid within subjective probability. The random variable, or, in De Finetti's terms, random quantity, $\|\chi\|$ now has infinitely many possible values in $[0, 1]$, rather than only the two possible values

$\{0, 1\}$ assumed by traditional probabilistic events. The value $[\![\chi]\!]$ is in De Finetti's framework now called a *prevision* of the random quantity $\|\chi\|$, but can still be defined by the above expression, and shares the relevant properties of probability.

To define this in greater detail, and to lend some concreteness to the above definition, let us consider traditional theorem proving and bag-of-words inference as special cases within this theory.

3.1 Special Case 1: Theorem Proving

Throughout our presentation, we have emphasized the fact that the degree of validity is a more fine-grained distinction of validity classes that generalizes over the traditional dichotomy between validity and non-validity.

We then only have to remark that, if we operate under the same assumptions as traditional deep inference, the same results will be achieved. More concretely, let us assume only two truth values, fully informative logical formulae, as well as complete and consistent theories of background knowledge. We then have $[\![\chi]\!]_{\mathcal{W}} = 1.0$ iff a traditonial theorm prover would prove χ.

We would expect this strategy, to work quite well for examples (1), (2.a), (2.b), (3.a), (3.b), (4) and (6), which require a great deal of informativity. Examples (2.c), (3.c), and (7), on the other hand, which would yield robustness, cannot be addressed with a traditional theorem prover.

3.2 Special Case 2: Bag-of-Words Inference

When we have a bag-of-words level of analysis for two pieces of text T and H, we can think of them in a logical representation as conjunctions, in which the atomic conjuncts are simply words, e.g.

$$\begin{array}{ll} & (\mathrm{T}) \quad \text{socrates} \wedge \text{is} \wedge \text{a} \wedge \text{man} \\ \hline \rightarrow & (\mathrm{H}) \quad \text{so} \wedge \text{every} \wedge \text{man} \wedge \text{is} \wedge \text{socrates} \end{array}.$$

Let's call this antecedent φ and the consequent ψ, and try to determine the degree of validity $[\![\varphi \rightarrow \psi]\!]$ for the bivalent case. This is possible using only some basic combinatorics.

Let Λ_φ be the set of propositional symbols, in this case words, appearing only in the antecedent, not in the consequent; $\Lambda_\varphi = \{\text{a}\}$. Similarly, let Λ_ψ be the set of propositional symbols appearing only in the consequent, not in the antecedent; $\Lambda_\psi = \{\text{so}, \text{every}\}$. Finally, let Λ_ω be the overlap, i.e. the set of propositional symbols appearing both in the antecedent and the consequent; $\Lambda_\omega = \{\text{socrates}, \text{is}, \text{man}\}$.

There are $\mathrm{N} = |\Lambda_\varphi \cup \Lambda_\psi \cup \Lambda_\omega| = 6$ atomic propositions altogether. We are dealing with the bivalent case, so there are $2^{\mathrm{N}} = 2^6 = 64$ possible valuations for this signature altogether. There are $2^{|\Lambda_\varphi|} = 2^1 = 2$ ways of assigning truth values to the antecedent, $2^{|\Lambda_\psi|} = 2^2 = 4$ ways of assigning truth values to the consequent, and $2^{|\Lambda_\omega|} = 2^3 = 8$ ways of assigning truth values to the overlap.

In order to make the implication "$\varphi \rightarrow \psi$" false, we must make the antecedent φ true, and the consequent ψ false. Clearly, only one out of the $2^{|\Lambda_\varphi \cup \Lambda_\omega|} = 2^1 * 2^3 = 16$ ways of assigning truth values to the antecedent makes the antecedent true. This is the case in which we assign the value 1 to all of the four conjuncts, thereby making the conjunction true. Out of the four conjuncts appearing in the consequent, this leaves two unassigned – we have already assigned truth values to the three conjuncts in the overlap set. There are 2^2 ways of assigning such truth values to the consequent, and only one of them makes the conjunction true, so the other $2^2 - 1 = 3$ all make the consequent false.

Therefore, out of the 2^6 possible valuations only $1 * 3 = 3$ valuations make the implication false. If we count zero for each of these three valuations, count one for all of the others, and divide the result by 2^6, we arrive at the value $[\![\varphi \rightarrow \psi]\!] = \frac{64-3}{64} = 0.95312$.

More generally,

$$[\![\varphi \rightarrow \psi]\!] = 1 - \frac{2^{|\Lambda_\psi|} - 1}{2^{|\Lambda_\psi| + |\Lambda_\phi| + |\Lambda_\omega|}}.$$

So we can express the degree of validity for a given candidate entailment in a closed form depending only on the forms of the words and how they match up against each other, assuming we encode a given piece of text simply as a conjunction in bivalent logic. Note that this closed form shares the same ordering properties as Dice's coefficient, the Jaccard index, or any other set overlap metric. These properties are as follows. (1) It acts as an overlap measure: Given φ or ψ, the ordering imposed by $[\![\varphi \rightarrow \psi]\!]$ on all ψ or φ, respectively, of the same length is the same as that imposed by $|\Lambda_\omega|$. (2) It performs length normalization: Given φ or ψ, the ordering imposed by $[\![\varphi \rightarrow \psi]\!]$ on all ψ or φ, respectively, is inverse to the length of such ψ or φ.

This strategy, would work well for examples (2.b), (2.c), (3.b.), (3.c), (4.a.), and (7), as these can easily be addressed using robust strategies. Examples (1), (2.a), (3.a), (4.b.), (4.c), and (6), on the other hand, require greater informativity, and cannot be addressed in this way.

Conclusions

Consider the following partition of buzzwords commonly used in NLP:

(a) shallow processing, robustness, probability, automatic acquisition, machine learning;

(b) deep processing, semantics, logic, knowledge engineering, artificial intelligence;

Furthermore, consider the following two statements overheard between NLP researchers in a pub:

(a) "...you must be very naive to believe you can reason about language in logic. Even if you could, you're missing the knowledge to prove things. Even if you had that, logic would still be too computationally complex."

(b) "…you must be rather ignorant to believe a machine learner will get you anywhere, if all you do is to feed it bags of words. It's just wrong from the point of view of logic, epistemology, linguistics, and whatever other theory you should care about."

Any attempt to describe, or put more meaningful labels on, these standpoints is quite unnecessary. Any member of the NLP community will be intimately familiar with the deeply entrenched paradigm which separates the field into (a) and (b).

To anyone subscribing to viewpoint (a), the overly restrictive consistency and completeness assumptions, as well as the theoretical limitations of the traditional notion of validity will seem like a bad idea. We agree. Probability theory can do better than that.

To anyone subscribing to viewpoint (b), the formula "every $\wedge$ man $\wedge$ is $\wedge$ socrates" will seem like a particularly bad idea, indeed. Again, we agree. Existing tools for semantic composition can do better than that. – But the following conclusions will perhaps come as more of a surprise.

In response to viewpoint (a), we can now say that knowledge and computational complexity are issues that are completely separate from the question of whether or not logic is a useful theoretic framework for approaching textual inference. It is all a question of how one represents text in logic. In the case of a bag-of-words representation, all the knowledge that is required is in the identities of the logical variables, and computational complexity is as little as that of evaluating a simple arithmetic expression.

In response to viewpoint (b), it is perhaps time to try accounting for the practical success of seemingly naive approaches like bag-of-words inference. Here, the robustness properties associated with the gradual notion of validity employed may be a key element.

In conclusion, we would like to emphasize, that our approach subscribes neither to viewpoint (a) nor to viewpoint (b) exclusively. Rather it is an attempt to make the two viewpoints complement, rather than contradict, each other within a single unified framework.

4 Monte Carlo Semantics

How do we approach the problem of computing $[\![\varphi \rightarrow \psi]\!]$ in all its generality? This problem is far from trivial, of course, because the decision problem for classical propositional logic can be reduced to it. If $\varphi \rightarrow \psi$ is a formula over Λ, then, using a naive approach, we could check whether $\varphi \rightarrow \psi$ is valid, i.e. whether $[\![\varphi \rightarrow \psi]\!] = 1.0$, by generating every possible valuation $w \in \mathcal{W}_{|\Lambda|,2}$. But there are $2^{|\Lambda|}$ such valuations.

Because we are considering, in the traditional case, the maximum or minimum truth value $\|\varphi \rightarrow \psi\|_w$ we can encounter for any w, this means we would have to run a model checker $2^{|\Lambda|}$ times, in the worst case.

Our approach is as follows: We exploit the fact that the arithmetic mean, in contrast to a maximum or a minimum, is very well behaved, when it comes to its statistically estimating it. We do not attempt to logically determine its exact value.

Thus, if $\mathrm{W} \subseteq \mathcal{W}$, we can use $[\![\varphi \rightarrow \psi]\!]_{\mathrm{W}}$ as an estimator for $[\![\varphi \rightarrow \psi]\!]_{\mathcal{W}}$. By statistical sampling theory, we know that the former will approach the latter as the sample size $|\mathrm{W}|$ approaches the population size $|\mathcal{W}|$. This sampling can be automated in a Monte Carlo method.

The central question that arises then, is how much information we obtain about $[\![\varphi \rightarrow \psi]\!]$ by simply assigning truth values to atomic propositions at random using a random number generator.

Let's consider a simple implication involving only atomic propositions: $[\![\mathrm{p} \rightarrow \mathrm{q}]\!] = 0.75$. We know that the truth table for this formula assigns the value 0 to only one valuation ($\|\mathrm{p}\|$ = 1, $\|\mathrm{q}\|$ = 0), and the value 1 to three valuations. Thus we have a $3/4$ chance of hitting the value 1.0 (error 0.25), and a $1/4$ chance of hitting the value 0 (error 0.75), which makes for a mean error of 0.375.

If we do this twice, we still have a $(3/4) * (3/4) = 9/16$ chance of hitting an average value of 1.0 (error 0.25), a $(3/4) * (1/4) + (1/4) * (3/4) = 6/16$ chance of hitting an average value of 0.5 (error 0.25) and finally a $(1/4) * (1/4) = 1/16$ chance of hitting an average of 0.0 (error 0.75). We have a mean error of 0.28125.

As we increase the number of trials, the mean error will decrease. But can we speed up the process? We can increase the number of truth classes. This is what a truth table for 3-valued Łukasiewicz logic would look like:

p	**1.0**	1.0	**1.0**	0.5	0.5	0.5	**0.0**	0.0	**0.0**
q	**1.0**	0.5	**0.0**	1.0	0.5	0.0	**1.0**	0.5	**0.0**
p → q	**1.0**	0.5	**0.0**	1.0	1.0	0.5	**1.0**	1.0	**1.0**

Four of these nine assignments coincide with bivalent logic, but we also insert five new values. We now have a mean truth value $[\![\mathrm{p} \rightarrow \mathrm{q}]\!] = 0.77$. We have a $6/9$ chance of hitting the value 1.0 (error 0.23), a $2/9$ chance of hitting the value 0.5 (error 0.27), and a $1/9$ chance of hitting the value 0.0 (error 0.77). The mean error is 0.296296.

We can now leave it as an exercise for the reader to verify that, if we repeat the process twice and obtain averages, we get an even smaller mean error of 0.19753.

We can increase the number of truth values in the logic, all the way to $\aleph_0$, where the value $[\![\mathrm{p} \rightarrow \mathrm{q}]\!]$, which is 1.0 iff $\mathrm{p} \leq \mathrm{q}$, takes on the value 1.0 only at a 0.5 chance. – Of course we can do this only in theory. In practice, any set of valuations, where the truth values come out of a random number generator for floating point numbers will fit into an M-valued logic for a large-enough M. The point is, that we do not want to restrict the possible values of truth values any further.

This raises some questions, as to whether the new logic, which now only proves a subset of the theorems provable in bivalent logic, is still a correct model of natural language semantics, and we believe it is. The example theorems listed in figure 1 can all be proved within such a logic.

φ :	Some elephants are intelligent,	φ :	$(e_1 \wedge i_1) \vee (e_2 \wedge i_2)$
ψ :	Some grey elephants are intelligent,	ψ :	$(e_1 \wedge g_1 \wedge i_1) \vee (e_2 \wedge g_2 \wedge i_2)$
χ :	Some clean grey elephants are intelligent.	χ :	$(e_1 \wedge c_1 \wedge g_1 \wedge i_1) \vee (e_2 \wedge c_1 \wedge g_2 \wedge i_2)$

	e_1	i_1	e_2	i_2	$(e_1 \wedge i_1) \vee (e_2 \wedge i_2)$	$= \varphi$	$\rightarrow \psi =$			$(\psi \rightarrow \varphi)$	χ	$(\varphi \rightarrow \chi)$	g_1	g_2	c_1	c_2
w_1	.99	.55	.47	.38	.55	.38	.55	.39	.84	1	.19	.64	.39	.19	.12	.97
w_2	.10	.58	.29	.00	.10	.00	.10	.10	1	1	.10	1	.98	.85	.62	.44
w_3	.13	.93	.59	.96	.13	.59	.59	.32	.73	1	.25	.66	.16	.32	.08	.25
w_4	.26	.64	.68	.74	.26	.68	.68	.68	1	1	.13	.45	.80	.99	.02	.13
w_5	.47	.10	.03	.76	.10	.03	.10	.10	1	1	.10	1	.65	.54	.10	.74
									.91	**1**		**.75**				

Figure 2: some of the examples in their logical representations

Example

To clarify the whole process, let us conclude this section by turning back to the elephant example, repeated here, together with a logical representation in figure 2.

Let the atomic proposition e_1 represent that individual 1 is an elephant, the proposition g_2, that individual 2 is grey, etc.

Now we can assign truth values to these propositions at random. Above, we have simply listed five different assignments of truth values. The truth values, listed in the first and last four columns, have not been carefully selected, but have simply been randomly generated, in this case, using a standard spreadsheet tool.

If we substitute the definition of the implication from definition 8 into defintion 5, it turns out that

$$\|\varphi \wedge \psi\| = \min(\|\varphi\|, \|\psi\|),$$
$$\|\varphi \vee \psi\| = \max(\|\varphi\|, \|\psi\|)$$

In the first valuation of the example, we have $\|e_1\|_{w_1} = 0.99$, $\|i_1\|_{i_1} = 0.55$, so $\|e_1 \wedge i_1\|_{w_1} = .55$, so "individual 1 is an intelligent elephant" is true to degree .55. Similarly, $\|e_2 \wedge i_2\|_{w_1} = .38$, so "individual 2 is an intelligent elephant" is true to degree .38. Finally $\|\varphi\|_{w_1} = \|(e_1 \wedge i_1) \vee (e_2 \wedge i_2)\|_{w_1} = \max(.55, .38) = .55$, so "some individual is an intelligent elephant" is true to a degree .55.

If we use the values $\|g_1\|$ and $\|g_2\|$, we can analogously determine $\|\psi\|_{w_1} = .39$. Since $1 - .55 + .39 = .84$, the implication stating "if some elephants are intelligent, then some grey elephants are intelligent" is true to a degree .84 in valuation w_1. The converse implication is true to a degree 1.0. Similarly, we can determine $\|\varphi \rightarrow \chi\|_{w_1} = .64$. Note that $\|\psi \rightarrow \varphi\| \geq \|\top\| = 1.0$, in accordance with (2.a), that $1.0 = \|\top\| < \|\varphi \rightarrow \psi\|$, in accordance with (2.b), and that $\|\varphi \rightarrow \psi\| < \|\varphi \rightarrow \chi\|$, in accordance with (2.c). It should be obvious, at this point, that this is not a coincidence.

While the fundamental logical properties are already fulfilled correctly, the exact truth values are still a function of the random valuation, we started out with. This is why, we now repeat the process for different valuations w_2, w_3, etc., to obtain mean truth values, i.e. degrees of validity, of $[\![\psi \rightarrow \varphi]\!] = 1.0$, and $[\![\varphi \rightarrow \psi]\!] = .91$, and $[\![\varphi \rightarrow \chi]\!] = .75$.

5 Initial Results & Future Directions

MCPIET is a direct implementation of this process and was tested successfully on section 1 of the FRACAS testsuite (Cooper et al., 1996). Now, we have also tried it on the RTE data for the first time, though with little success visible in the form of statistics over this dataset. This is why we have emphasized before, that our main contribution currently lies with our theory.

The reason why we have seen little practical success is not because of a flaw in the inference mechanism, but because of the naive setup of the frontend infrastructure that converts text into logical formulae.

MCPIET can use either BOXER and the C&C TOOLS (Curran et al., 2007) or the ERG (Copestake and Flickinger, 2000) for this translation. In the case of the ERG, we first have to scope the text. For the sake of simplicity, we currently always choose the top ERG parse and scope the quantifiers in the order in which they appear in the text. We then apply some reductions to translate generalized quantifiers into first-order structures. In the case of the C&C TOOLS, no parses other than the top one are available to begin with. Semantic composition, scope selection and first-order reduction are then done by BOXER.

The main problem here was that we effectively force the grammar into venturing a wild guess as to which in a number of ambiguous readings to assign to a given sentence. So we are not yet implementing the strategy suggested by our theory of falling back to less informative representations, where information is unavailable or syntactically unknown. Rather, we have wrong information represented in our logical formulae, which is obviously counterproductive. To solve this problem, we are currently investigating methods of removing this noise by generalizing logically over the semantic representations of multiple parses.

Another main shortcoming of the current implementation is

that we do not yet support any kind of coreference resolution or other semantically relevant discourse phenomena, not even anaphora. We have not implemented special recognition and reasoning with dates, times, geographical locations, organization names, etc.

6 Concluding Remarks

For all these reasons outlined in the previous section, it comes as little surprise that MCPIET does not perform competitively yet with other systems in an RTE evaluation. While many other systems are already pushing the limits of what is theoretically possible within their approaches, development on MCPIET has hardly even begun. It is these theoretical limits, which have been our central concern herein.

More particularly, we were concerned with the design goals of informativity and robustness. We have shown a catalogue of example inferences, to define exactly what properties one would expect of a system, in connection with a claim to informativity or robustness.

We have situated current approaches to deep and shallow inference within our theoretic framework, and shown their limits by characterizing the informativity/robustness tradeoff, a well-known experience to the community, yet a concept which is usually not quite trivial to pin down theoretically.

Our theoretical framework relies on a notion of a *degree of validity*, and we have provided some evidence to suggest that precisely this concept of graded validity may play a crucial role in the robustness properties of shallow inference. At the same time our framework still supports informationally rich semantic representations and background theories, which play a crucial role in the informativity properties of deep inference.

From this point of view, we have then posed the problem of deep/shallow integration within our new theory, and also proposed a solution to it we have called *Monte Carlo Semantics*.

Acknowledgments

I would like to thank Ann Copestake and Ulrich Bodenhofer for their continued support throughout this research project. I have been supported financially by an EPSRC studentship, a Cambridge European Bursary, and a DOC-fellowship by the Austrian Academy of Sciences, and would like to thank the benefactors who made this possible.

References

Johan Bos and Katja Markert. 2005. Combining shallow and deep nlp methods for recognizing textual entailment. In Ido Dagan, Oren Glickman, and Bernardo Magnini, editors, *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (RTE-1)*.

Johan Bos and Katja Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment (RTE-2)*.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning. 2007. Learning alignments and leveraging natural logic. In *Proceedings of the Workshop on Textual Entailment and paraphrasing (RTE-3)*.

C. C. Chang. 1958. Proof of an axiom of lukasiewicz. *Transactions of the American Mathematical Society*, 87(1):pp. 55–56, January.

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. Technical Report D16, FraCaS project deliverable, January.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second Linguistic Resources and Evaluation Conference*, pages 591–600, Athens, Greece.

James R Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the Demonstrations Session of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*.

Bruno De Finetti. 1974. *Theory of probability : a critical introductory treatment*. Wiley, London. Translation of Teoria delle probabilita.

Jan Łukasiewicz and Alfred Tarski. 1930. Untersuchungen über den aussagenkalkül. *Comptes rendus des séances de la Société des Sciences et des Lettes de Varsovie*, 23:39–50, March.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the Workshop on Textual Entailment and paraphrasing (RTE-3)*.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling-08)*.

C. A. Meredith. 1958. The dependence of an axiom of lukasiewicz. *Transactions of the American Mathematical Society*, 87(1):p. 54, January.

Alan Rose and J. Barkley Rosser. 1958. Fragments of many-valued statement calculi. *Transactions of the American Mathematical Society*, 87(1):pp. 1–53, January.

Notes

[1] The fragment of the model theory considered here consisting of implication and negation as basic operators as well as the attached notion of validity were first introduced in a publication by (Łukasiewicz and Tarski, 1930), but are correctly attributed to Jan Łukasiewicz alone.

[2] The full set of operators considered here were used by (Rose and Rosser, 1958) and now appear throughout the relevant literature on multi-valued logic.

[3] The axioms were given by Łukasiewicz himself (Łukasiewicz and Tarski, 1930), in addition to a fifth axiom. He conjectured that these five theses would form an axiomatization for the semantic system he was considering, but did not present a completeness proof to that extent. In 1935, M. Wajsberg claimed to have proved this completeness result, but such a proof never appeared in print. A completeness result for these five axioms was established by (Rose and Rosser, 1958). At the same time, it was also found that the fifth axiom considered by Łukasiewicz was in fact dependent (Meredith, 1958; Chang, 1958).