On Quantitative Similarity-based Semantics for Counterfactuals

Marco Cerami Pere Pardo

Artificial Intelligence Research Institute (IIIA - CSIC)
Barcelona (Spain)

September 23th 2009, Liblice
WUPES'09

Lewis' System of Spheres for Classical Counterfactuals

As commonly understood, a *counterfactual* is a conditional sentence whose antecedent is false.
Material implication (from e.g. classical logic) makes conditional sentences with a false antecedent unconditionally true.
In contrast with material implication sentences, a false antecedent does not make the counterfactual conditional automatically true.
Lewis (1973) provided a semantics for classical logic counterfactuals based on a type of structures called *Systems of Spheres*.

Definition

A *system of spheres* on a set W, denoted by \$, is a subset of $\mathcal{P}(W)$ such that \$ is

1. totally ordered by inclusion, and
2. closed under arbitrary unions and non-empty intersections.

Lewis' Semantics for Classical Counterfactuals

Let $[\varphi] \subseteq W$ denote the set of possible worlds in which φ is true.

Definition

Let W be a set of possible worlds, $\$ a system of spheres on W, $i \in W$ and, for every formula φ, then:

A counterfactual $\varphi \boxdot\!\!\rightarrow \psi$ is true in a world i, iff

1. either $[\varphi] = \emptyset$
2. or $w \models \psi$, for every $w \in \{[\varphi] \cap \bigcap\{S \in \$^i | S \cap [\varphi] \neq \emptyset\}\}$.

A counterfactual $\varphi \Diamond\!\!\rightarrow \psi$ is true in a world i, iff

1. either $[\varphi] = \emptyset$
2. or $w \models \psi$, for some $w \in \{[\varphi] \cap \bigcap\{S \in \$^i | S \cap [\varphi] \neq \emptyset\}\}$.

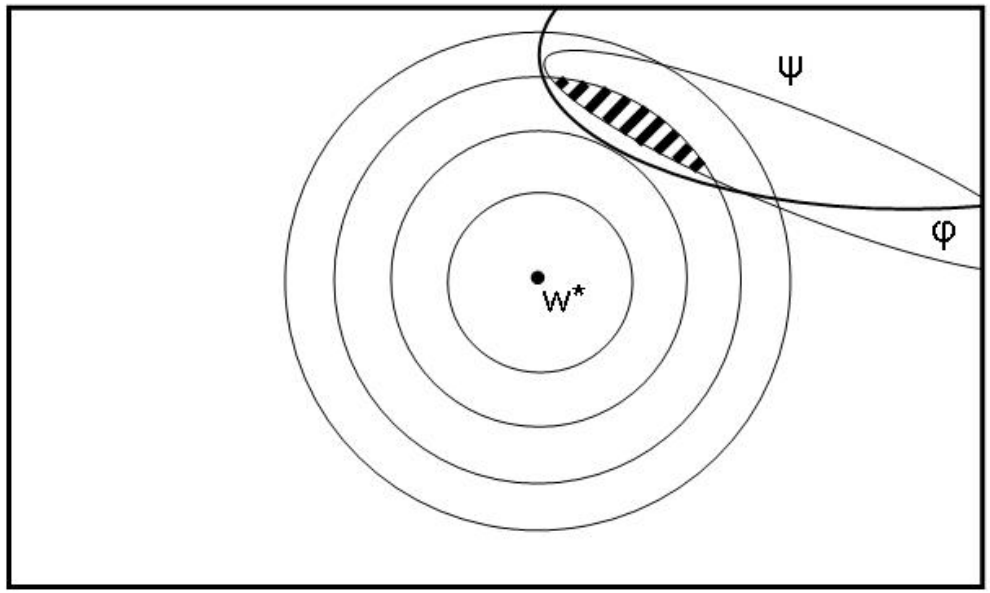


Figure: Semantics for *would*-counterfactuals

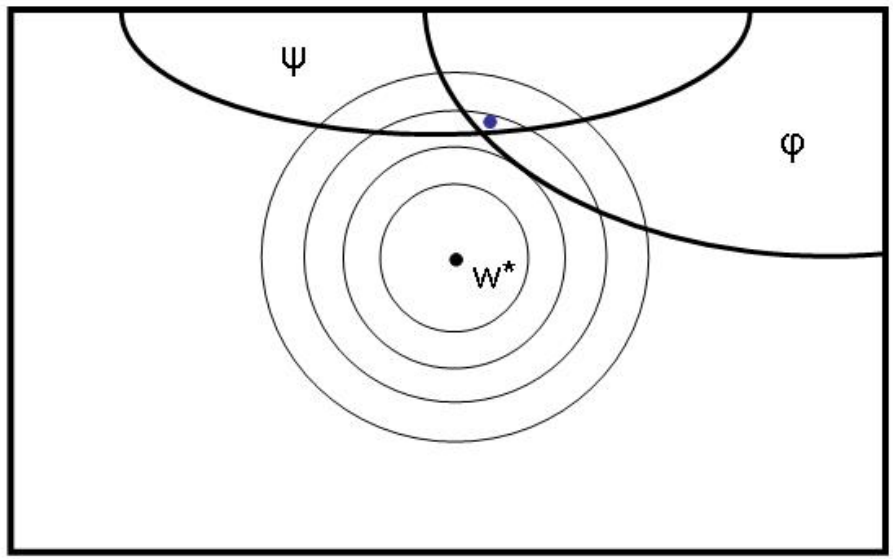


Figure: Semantics for *might*-counterfactuals

Would-might interdefinability in the classical case.

In the classical framework, like in Modal Logic, it is usually sufficient to assume as primitive only one of the above symbols and to define the other by means of negation.

Proposition

Let φ, ψ be arbitrary formulas, then:

1. $\varphi \mathrel{\Box\!\!\rightarrow} \psi \equiv \neg(\varphi \mathrel{\Diamond\!\!\rightarrow} \neg\psi)$
2. $\varphi \mathrel{\Diamond\!\!\rightarrow} \psi \equiv \neg(\varphi \mathrel{\Box\!\!\rightarrow} \neg\psi)$

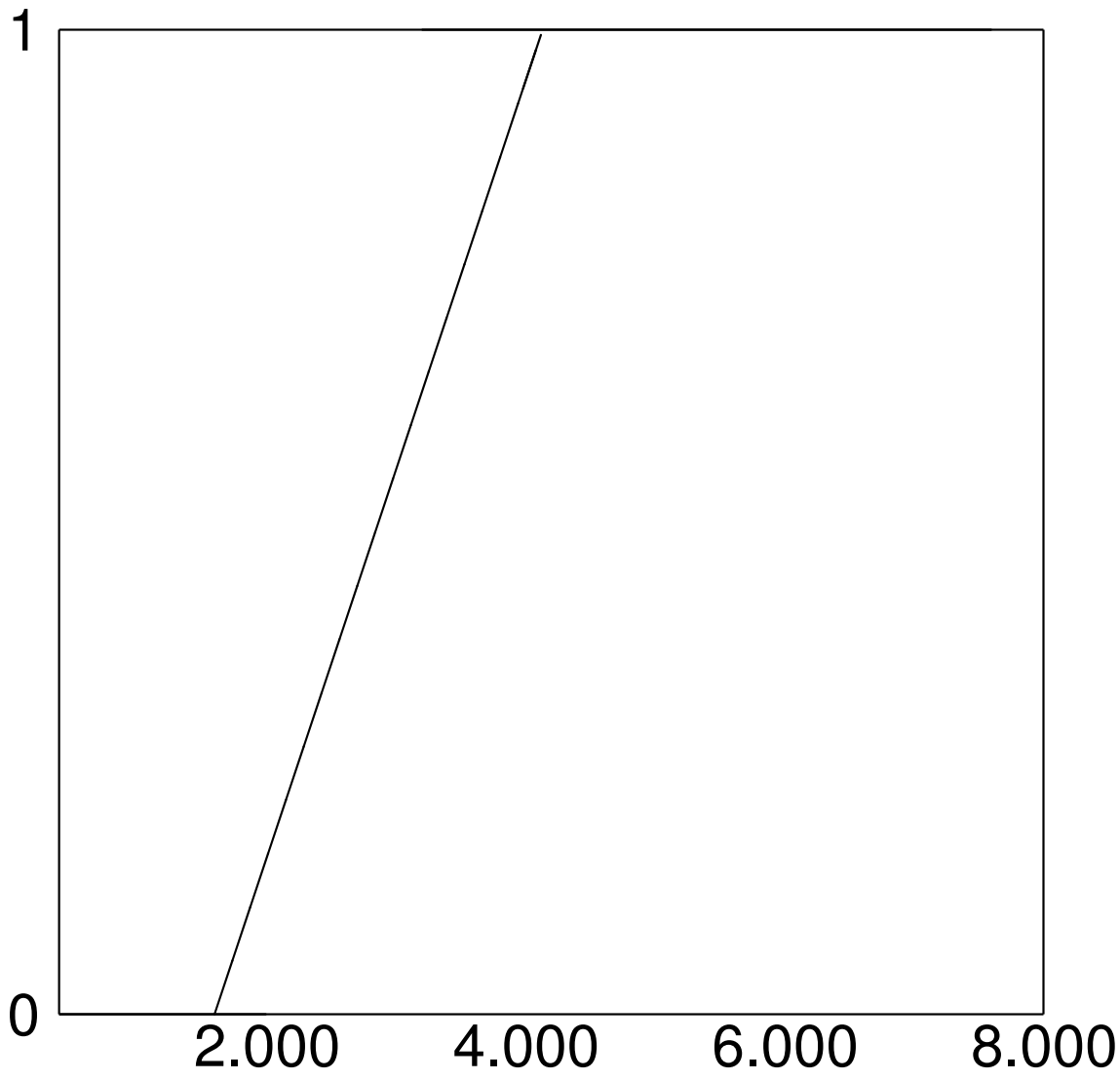
Vague Sentences

We consider a *vague sentence* as a sentence that, by the nature of the meaning involved cannot be understood as merely true or false.
As an example, if we fix that a tall man is a man whose height is greater or equal to 1.80 m, we cannot consider a man who is 1.79 m tall as a short man, even if he is not tall. Considerations like these drove to the ancient Sorites paradox; a modern way to overcome such paradox has been to consider *fuzzy sets*.
Zadeh (1965) defined a fuzzy set M as a set whose characteristic function[1] χ_M, is a function which returns a real value between 0 and 1, i.e. $\chi_M(x) \in [0, 1]$. Intuitively, if M is the set of tall men and x is a man who is 1,79 m tall, then we may have, say, $\chi_M(x) = 0.95$.

[1] A *characteristic function*, in the classical framework, is a function χ such that $\chi_M(x) = 1$ if an individual x is element of a set M and $\chi_M(x) = 0$ otherwise.

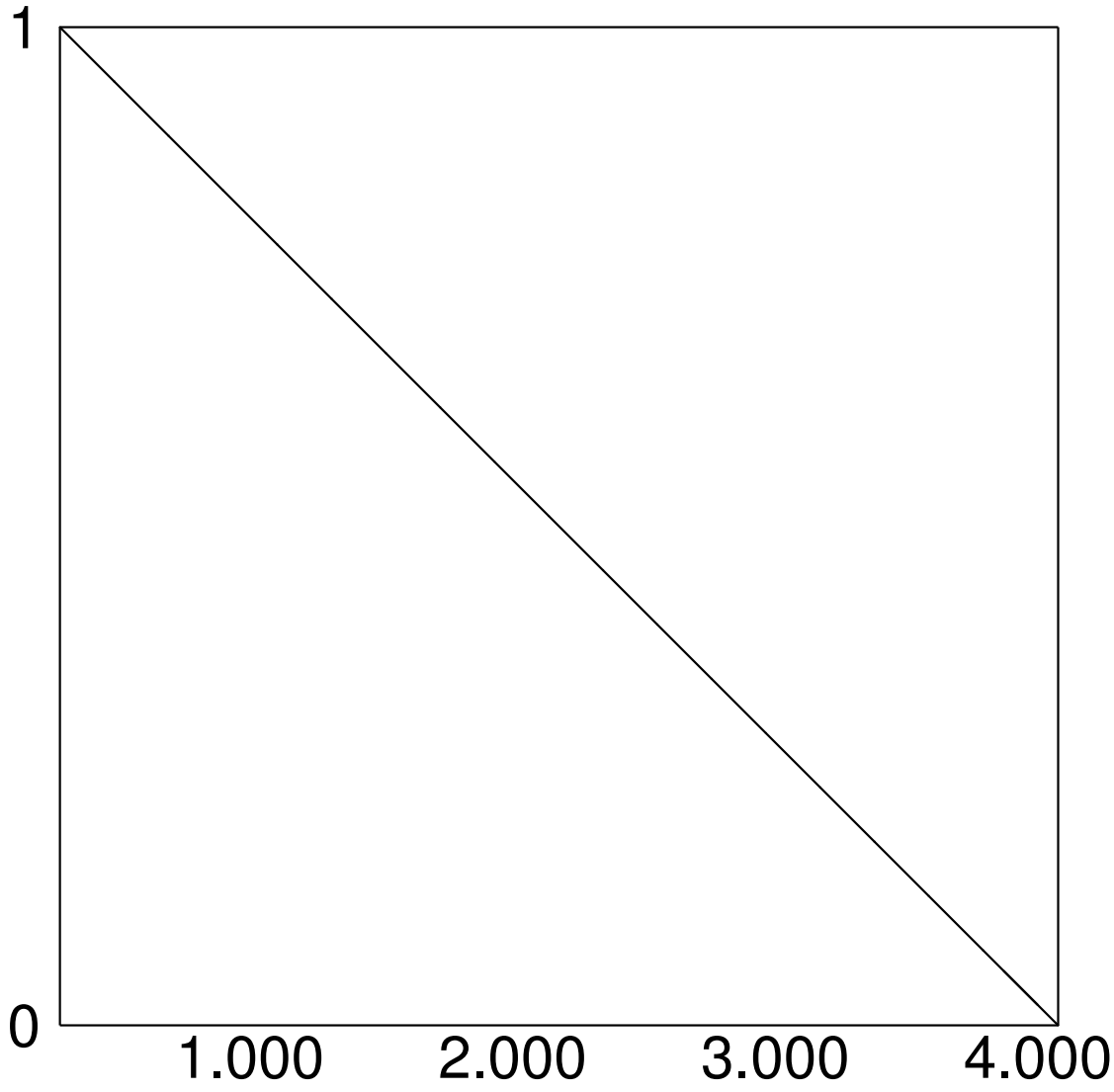
Example of vague sentences (i)

The mountain x is height.



Example of vague sentences (ii)

The place *x* is near from Prague.



t-norms

Definition

A *t-norm* is a binary operation $* : [0,1]^2 \to [0,1]$ such that:

1. $*$ is commutative and associative.
2. $*$ is non-decreasing in both arguments.
3. For every $x \in [0,1]$, it holds that $1 * x = x$ and $0 * x = 0$.

If, moreover, $*$ is a continuous mapping from $[0,1]^2$ to $[0,1]$, we talk about a *continuous* t-norm.

The main examples of continuous t-norms are:

1. *Łukasiewicz* t-norm (denoted by Ł), defined by the function: $x * y = max(0, x + y - 1)$,
2. *Gödel* t-norm (denoted by G), defined by the function: $x * y = min(x, y)$,
3. *Product* t-norm (denoted by Π), defined by the function: $x * y = x \cdot y$.

Residua

Definition

Let $*$ be a t-norm, then its residuum is a binary operation $\Rightarrow_* : [0,1]^2 \to [0,1]$ such that, for every $x, y \in [0,1]$:

$$x \Rightarrow_* y = sup\{z \in [0,1] \mid x * z \leq y\}$$

Intuitively, residua are the semantics of implications, indeed, it the framework of Fuzzy Logic, the expression *material implication* is substituted by the more general *residuated implication*.

The main examples of residua are:

1. *Łukasiewicz*, defined by the function: $x \Rightarrow_* y = min(1, 1 - x + y)$,
2. *Gödel*, defined by the function: $x \Rightarrow_* y = 1$, if $x \leq y$, y, otherwise,
3. *Product*, defined by the function: $x \Rightarrow_* y = min(1, \frac{y}{x})$.

t-norm based logics

Definition

Let $* \in \{Ł, G, \Pi\}$, then a *$*$-based propositional fuzzy logic* (denoted by L_*) is the least set of sentences which includes the axioms of such logic and is closed under Modus Ponens.

Definition

Let φ, ψ formulas, for each t-norm $* \in \{Ł, G, \Pi\}$, we inductively define the *propositional evaluation* from a morphism $e : \mathrm{Var} \to [0, 1]$ as follows:

1. $e(\bot) = 0$,
2. $e(\top) = 1$,
3. $e(\varphi \wedge_* \psi) = e(\varphi) * e(\psi)$,
4. $e(\varphi \to_* \psi) = e(\varphi) \Rightarrow_* e(\psi)$.

Quantitative vs meta-linguistic criteria

In a classical framework, it is the same to say that a given sentence φ *is true*, *holds* or that it has value 1.
So, in this case, it is possible to express the truth value of a sentence by means of a meta-linguistic expression.
For this reason, Lewis does not consider important to give a quantitative definition of his truth conditions for counterfactuals.
In a context of multi-valued sentences, we deal with a different situation: as a simple remark, to give meta-linguistic account for each truth value, we would need an infinite set of adjectives ranging between *true* and *false*...
and we are not sure that there exists any natural language that posses all these expressions.
The natural choice seems then to be using numbers.

Distance based on a single sentence

Any pair of worlds (or models) must differ in at least a sentence. This sentence induces a distance function between worlds. The starting point of our suggested formalism consists in this notion of distance.

Definition

Let $\varphi \in Fm$ and W a set of worlds with $w, w' \in W$. We define a *distance function* on W as

$$d_\varphi(w, w') = |e_w(\varphi) - e_{w'}(\varphi)|$$

Var-based Distance

Two worlds w, w' will in general differ in more than one sentence. The difference between w and w' reduces to that between atomic formulas (in Var) according to the previous definition.
A simple account of this difference is given by the sum of the distances of evaluations of propositional variables which differ between two given worlds.

Definition

Let Var be the set of propositional variables and W a set of worlds with $w, w' \in W$. We define a Var-based *distance function* on W as

$$d_{\text{Var}}(w, w') = \Sigma_{p \in \text{Var}} d_p(w, w')$$

where d_p is the distance function previously defined.

A quantitative Similarity measure

Our aim is, indeed, to define a similarity measure between pairs of worlds. This can be simply obtained from the Var-based distance function defined above.

Definition

Let Var be the set of propositional variables and W a set of worlds with $w, w' \in W$. We define a *similarity measure* on W as

$$\mathcal{S}(w, w') = \frac{1}{1 + d_{\text{Var}}(w, w')}$$

Observe that $\mathcal{S}(w, w) = 1$

Systems of Spheres based on Quantitative Similarity

Finally, we can define a semantics that is based on the above defined quantitative Similarity measure $\mathcal{S}$.

Definition

Let W be a set of possible world and $r \in [0,1]$, then we denote by S^r the sphere:

$$S^r = \{w \in W \mid \mathcal{S}(w^\star, w) \geq r\}$$

Now,

$$\$^{[0,1]} = \{S^r \mid r \in [0,1]\}$$

It is straightforward that the above definition fits with Lewis' definition of a System of Spheres.

Vague counterfactuals

By a *vague counterfactual* we understand a counterfactual involving vague sentences, i.e., sentences, that are not merely true or false, but can be evaluated in $[0, 1]$.
This implies that the counterfactual, as a formula, can be evaluated in $[0, 1]$ as well.
The most widely accepted definition of a (classical) counterfactual is to be a conditional with a false (within the actual world) antecedent, but, while in the classical framework there is no difference between speaking about a *false* antecedent and about an antecedent *that is not true*, in a multi-valued framework we have the chance of smarter specifications which allow a wider expressivity.

Negated antecedents

As an example, consider the sentence:
”If I was tall, I would reach for the roof”
This sentences assumes that in the actual world, I am not tall in degree 1, not that I am tall in degree 0; in other words, I can think that I’m not tall even without thinking that I’m short.
A first choice would be simply to consider worlds where the antecedent φ takes value 1; more generally, we could consider worlds where φ takes a value higher than, say, r; a particular case of the latter would consist in setting r to be the actual value for φ. In the following we consider these accounts and present a possible way to formalize them.

Definition of 1-semantics

We define first a simple extension of the semantics for the crisp case, now defining φ-worlds as worlds w where φ is 1-true: $e^{\$}_w(\varphi) = 1$.

Definition

The 1-semantics of *would* counterfactuals $\varphi \boxdot\!\!\rightarrow \psi$ and *might* $\varphi \Diamond\!\!\rightarrow \psi$ is defined by:

$$
\begin{array}{lcll}
e^1_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) & = & \inf\{ & e_w(\psi) \mid e_w(\varphi) = 1 \text{ and } \\
 & & & w \in \bigcap\{S \in \$: \exists w' \in S(e_{w'}(\varphi) = 1)\}\} \\
e^1_{w^\star}(\varphi \Diamond\!\!\rightarrow \psi) & = & \sup\{ & e_w(\psi) \mid e_w(\varphi) = 1 \text{ and } \\
 & & & w \in \bigcap\{S \in \$: \exists w' \in S(e_{w'}(\varphi) = 1)\}\}
\end{array}
$$

Reduction of the classical framework to ours

The restriction to crisp systems makes the inf- and $\forall$- clauses (and, hence, the respective semantics) equivalent.
In other words, it is possible to prove that Lewis' semantics for counterfactuals is a particular case of the preceding semantics under the condition that the evaluations are restricted to $\{0, 1\}$, as in classical logic.

Proposition

Let $e^c_w(\cdot)$ denote Lewis' semantics and W a set of classical possible worlds (i.e. $W \subseteq \{e^c_w : \mathrm{Var} \to \{0,1\}\}$). For any classical system of spheres $\$$ and any world $w \in W$, the 1-semantics $e^1_w(\cdot)$ definition gives:

$$e^c_w(\varphi \boxdot\!\!\rightarrow \psi) = e^1_w(\varphi \boxdot\!\!\rightarrow \psi)$$

and

$$e^c_w(\varphi \diamond\!\!\rightarrow \psi) = e^1_w(\varphi \diamond\!\!\rightarrow \psi)$$

Interdefinability

In the particular case of Łukasiewicz, we also have that classical inter-definability of *would* and *might* counterfactuals is preserved (this is due to the fact that $*_{\mathbf{Ł}}$ is the only t-norm whose negation $e(\varphi) \Rightarrow \overline{0}$ is involutive, and hence behaves well with inf).

In the next result, in Ł logic, we assume non-counterfactual connectives are evaluated according to Łukasiewicz semantics:
$e(\varphi * \psi) = \max\{0, e(\varphi) + e(\psi) - 1\}$, and
$e(\varphi \to_{\mathbf{Ł}} \psi) = \min\{1, 1 - e(\varphi) + e(\psi)\}$.

Proposition

Let W be a set of possible worlds, $w \in W$ and let $\neg$ denote Łukasiewicz negation, then:

$$e^1_w(\varphi \diamond\!\!\rightarrow \psi) = e^1_w(\neg(\varphi \Box\!\!\rightarrow \neg\psi))$$

and

$$e^1_w(\varphi \Box\!\!\rightarrow \psi) = e^1_w(\neg(\varphi \diamond\!\!\rightarrow \neg\psi))$$

A troubling example

It is indeed possible to further generalize such a semantics to the $> r$-case, in order to obtain a more refined tool with respect to vague counterfactual.
The task, however, seems to be non-trivial: observe first that a simple (further) generalization of the previous 1-semantics to the $> r$-case will have the counterintuitive result of making $\varphi \boxdot\!\!\rightarrow \varphi$ non-tautological (i.e. not 1-true).

Example

For a fixed $r < 1$, replace (in the 1-semantics above) in each $e_{(\cdot)}(\varphi) = 1$ clause, the expression $= 1$ by expression $\geq r$. Also assume the actual world assigns φ a degree $< r$. Now, say the \$-system contains a φ-world w (i.e. $e^{\$}_w(\varphi) \geq r$) in the closest sphere S. Then, according to this semantics, we would have $e^{\geq r}_{w^\star}(\varphi) = \inf\{e_w(\psi) \mid e_w(\varphi) \geq r \text{ and } w \in \bigcap\{S \in \$: \exists w' \in S(e_{w'}(\varphi) \geq r)\}\}\} = r < 1$. Hence $\varphi \boxdot\!\!\rightarrow \varphi$ is not a tautology.

Definition of $\geq r$-semantics

A simple way to overcome such a problem, inspired on Modal Logic, is to define the value of the counterfactual from the value of the corresponding residuated implication in the possible world w, such that $e_w(\varphi) \geq r$ and w is in the closer sphere to the actual world.
A formal account of this condition is given in the following definition:

Definition

For a given $r \in [0,1]0$, let

$$\mathbb{K}^r = \{w \in W : e_w(\varphi) \geq r \text{ and } w \in \bigcap \{S \in \$ \mid \exists w' \in S(e_{w'}(\varphi) \geq r)\}\}$$

Then we define the $\geq r$-semantics of $\boxdot\!\!\rightarrow$ and $\diamond\!\!\rightarrow$ as follows:

$$\begin{aligned} e^{\geq r}_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) &= \inf\{\, e_w(\varphi \rightarrow_* \psi) \mid w \in \mathbb{K}^r\} \\ e^{\geq r}_{w^\star}(\varphi \diamond\!\!\rightarrow \psi) &= \sup\{\, e_w(\varphi * \psi) \mid w \in \mathbb{K}^r\} \end{aligned}$$

Intuitive interpretation

Intuitively, this kind of implication is appropriate when we can give an exact degree (lower bound) to the expected value of the antecedent, as in the following example:

Example

Assume we interpret truth at degree (at least) 0.8 as *very* true (correspondingly, *very [humanely] tall*, *very [mountainly] tall*, ...). Then the following sentences:

(1) If I was very rich, I would be happy. $(\varphi, 0.8) \boxdot\!\!\rightarrow \psi$

(2) If I was very rich I might be happy. $(\varphi, 0.8) \Diamond\!\!\rightarrow \psi$

Reduction of Lewis' semantics to ours

This last kind of definition, does not violate Lewis' truth condition in the classical case, but it is a generalization of it. Indeed, if we restrict the truth values of the sentences involved in the counterfactual to $\{0, 1\}$, we obtain Lewis' truth conditions:

Proposition

Setting $r = 1$ for a given system of spheres $ and world $w^\star$, $\geq r$-semantics (in classical valuations only) collapses to Lewis'.

$$\begin{aligned} e_{w^\star}(\varphi \mathrel{\Box\!\!\rightarrow} \psi) &= e^{r=1}_{w^\star}(\varphi \mathrel{\Box\!\!\rightarrow} \psi) \\ e_{w^\star}(\varphi \mathrel{\Diamond\!\!\rightarrow} \psi) &= e^{r=1}_{w^\star}(\varphi \mathrel{\Diamond\!\!\rightarrow} \psi) \end{aligned}$$

Interdefinability

In this case, too, Łukasiewicz semantics allows us to obtain an interdefinability result of the kind of Lewis.

Proposition

Let W be a set of possible worlds, $w \in W$ and let $\neg$ denote Łukasiewicz negation, then:

$$e_w^{\geq r}(\varphi \diamond\!\!\rightarrow \psi) = e_w^{\geq r}(\neg(\varphi \Box\!\!\rightarrow \neg\psi))$$

and

$$e_w^{\geq r}(\varphi \Box\!\!\rightarrow \psi) = e_w^{\geq r}(\neg(\varphi \diamond\!\!\rightarrow \neg\psi))$$

Definition of more-than-actual Semantics

The main goal of a semantics for vague counterfactual is, however, to define the truth value of a counterfactual in the case when the antecedent has a value greater than the actual one. Taking as a starting point the last definition, the weakest condition to be imposed on antecedents of vague counterfactuals will consider worlds where its truth-value is (at least) just slightly higher than in the actual world.

Definition

Let W be a set of possible worlds, $w^\star \in W$ and \$ a system of spheres, then:

$$(1)\ e^{>w^\star -}_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) = inf_{r>e_{w^\star}(\varphi)}\{e^r_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi)\}$$

$$(2)\ e^{>w^\star +}_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) = sup_{r>e_{w^\star}(\varphi)}\{e^r_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi)\}$$

Intuitive interpretation (i)

As we have seen above, starting from the same $\geq$ r-semantics, we can obtain two different kind of definition.
By means of such a difference it is possible to give account of different natural language counterfactuals.
By (1) it is possible to give truth conditions for counterfactuals with antecedents that, in the possible world considered, have a value just higher than the actual one.
Some example can be:
If Prague was nearer, there would be a tube line to arrive there.
or
No matter how much richer I was richer, I would not work.
However, even if this seems to be the proper semantics for vague would counterfactual, it make sense with a finite set of truth values, since, with a dense one, the value of the antecedent seems to collapse to the actual one.

An example

To show that, in finitely-valued fuzzy logics, the $> w^\star$-semantics is a particular case of $\geq r$-semantics consider:

$$e^{>w^\star}_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) = inf\{\, e_w(\varphi \rightarrow \psi) \mid e_w(\varphi) \geq e_{w^\star}(\varphi)^+\}, \text{ and}$$

$$w \in \bigcap \{S \in \$: \exists w' \in S(e_{w'}(\varphi) \geq e_{w^\star}(\varphi)^+)\}\}$$

The more-than-actual semantics in the dense-valued case, typically $[0,1] \cap \mathbb{Q}$, lead to a different situation, since, this time, definition (1) is not a particular case of the $\geq r$-semantics. To see this fact, let φ, ψ be two vague sentences such that, for each possible world $w \in W$, $e_w(\psi) = 1 - e_w(\varphi)$ and suppose that, for each $S, S' \in \$$, it holds that if $S \subseteq S'$, then there exist $w \in S$ and $w' \in S'$, such that $e_w(\varphi) < e_{w'}(\varphi)$.
So, if we consider the counterfactual $\varphi \boxdot\!\!\rightarrow \psi$, we have that, for each $r \geq (e_{w^\star}(\varphi)$, $e^r_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) > e^{>w^\star -}_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi)$.

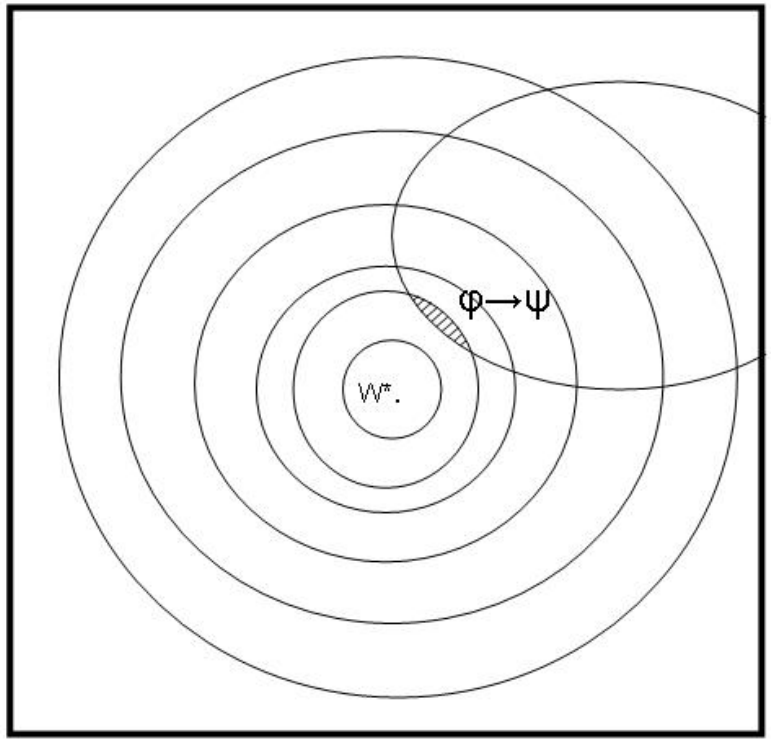


Figure: $\geq$ r-semantics when $e_w(\varphi) > e_{w^\star}(\varphi)$

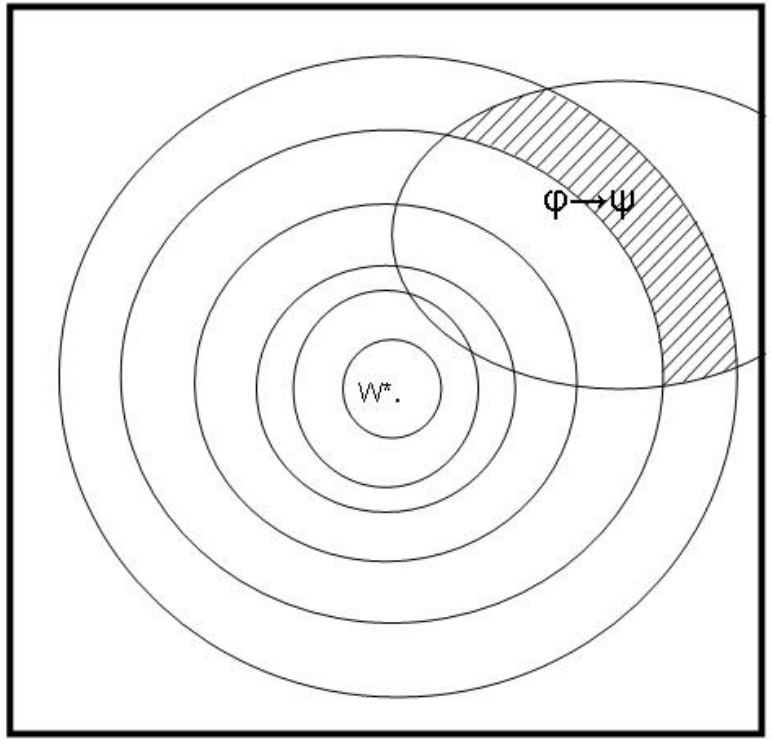


Figure: More-than-actual semantics

Intuitive interpretation (ii)

By (2) it is possible to give truth conditions for counterfactuals with antecedents that, in the possible world considered, have a value enough higher than the actual one.
Some example can be:
If Prague was near enough, we would come each weekend.
or
If I was rich enough, I would be happy.
This settling works also with a dense set of truth values, but is indeed too narrow to give an account for each kind of counterfactual.

Dual counterfactuals

In the Łukasiewicz framework we have that the dual of the above definitions are:

$$(1)\ e^{>w^\star -}_{w^\star}(\varphi \diamond\!\!\rightarrow \psi) = sup_{r>e_{w^\star}(\varphi)}\{e^r_{w^\star}(\varphi \diamond\!\!\rightarrow \psi)\}$$
$$(2)\ e^{>w^\star +}_{w^\star}(\varphi \diamond\!\!\rightarrow \psi) = inf_{r>e_{w^\star}(\varphi)}\{e^r_{w^\star}(\varphi \diamond\!\!\rightarrow \psi)\}$$

Some example can be of (1):
If Prague was near enough, we might come each weekend.
or
If I was rich enough, I might be happy.
While some example of (2) can be:
If Prague was nearer, there might be a tube line to arrive there.
or
No matter how much richer I was richer, I might not work.

Quantitative Similarity-based Semantics

The quantitative similarity-based semantics can give us an alternative way to define truth conditions for vague (as well as classical) counterfactuals.
As an example, we will report here only the case of the $\geq r$-semantics, the other cases follow directly from this definition.

Definition

Let W be a set of possible worlds, $w^\star$ the actual world and $\mathcal{S}$ the Similarity measure defined before, then, for each pair of formulas φ, ψ:

$$e^{\geq r}_{w^\star}(\varphi \boxdot\!\!\rightarrow \psi) = \quad \inf_{w \in W}\{ \quad e_w(\varphi) \geq r * \; inf_{w' \in W}\{\mathcal{S}(w^\star, w') > \mathcal{S}(w^\star, w) \Rightarrow e_{w'}(\varphi) \Rightarrow e_w(\varphi \rightarrow \psi)\}$$

$$e^{1}_{w^\star}(\varphi \diamond\!\!\rightarrow \psi) = \quad \sup_{w \in W}\{ \quad e_w(\varphi) \geq r * \; inf_{w' \in W}\{\mathcal{S}(w^\star, w') > \mathcal{S}(w^\star, w) \Rightarrow e_{w'}(\varphi) * e_w(\varphi * \psi)\}$$

Equivalence result

It is possible to prove that this last definition is equivalent to the metalinguistic one defined in a previous section.
Indeed, this is equivalent to show the following simpler result.

Proposition

Let W be a set of possible worlds, $\mathbb{K}^r$ as defined before, and $\$$ a system of spheres based on $\mathcal{S}$, then, for every formula φ and $w \in W$:

$$w \in \mathbb{K}^r \iff$$

$$e_w(\varphi) \geq r \ * \ inf_{w' \in W}\{\mathcal{S}(w^\star, w') > \mathcal{S}(w^\star, w) \Rightarrow_* e_{w'}(\varphi) < r\} = 1$$

Equivalence result

It is possible to prove that this last definition is equivalent to the metalinguistic one defined in a previous section.
Indeed, this is equivalent to show the following simpler result.

Proposition

Let W be a set of possible worlds, $\mathbb{K}^r$ as defined before, and $\$$ a system of spheres based on $\mathcal{S}$, then, for every formula φ and $w \in W$:

$$w \in \mathbb{K}^r \iff$$

$$e_w(\varphi) \geq r \ * \ inf_{w' \in W}\{\mathcal{S}(w^\star, w') > \mathcal{S}(w^\star, w) \Rightarrow_* e_{w'}(\varphi) < r\} = 1$$

Interdefinability result and the fact that this quantitative similarity-based definition is a generalization of Lewis' one for the crisp case, follow easily from the last proposition.