A Computational Approach to Conceptual Blending within Counterfactuals

Mark Lee & John Barnden
School of Computer Science
University of Birmingham

Abstract.
An important topic in philosophy and the semantics of natural language has been the representation and verification of counterfactuals. The majority of approaches have taken a classical view of language where statements can be reasoned about solely in terms of their truth functions. This has led to several well known problems in the verification of counterfactuals. We take another view, inspired by cognitive linguistics, that language relies on the mappings between the conceptual structures (or domains) so that phenomena such as metaphor, blending and analogy are not exceptional uses of language but rather central aspects. To this end, we present an analysis of counterfactuals based on mapping between different reasoning spaces and the drawing of analogies between these spaces. This analysis is implemented in an artificial intelligence reasoning system.

The system, called ATT-Meta, was originally designed for uncertain metaphor-based reasoning (and belief reasoning), and was applied without modification to the task of verification of counterfactuals. This was possible because the system approaches metaphor by a pretence mechanism akin to counterfactual reasoning, and because it already contained powerful mechanisms for resolving conflicts between competing arguments.

1 Introduction

The interpretation of counter-factual expressions has been a key topic in philosophy, linguistics and semantics. A typical counter-factual has the form:

(1) "If John had sold his shares then he would have made a profit."

For the statement to be *counter*-factual, both the antecedent and the conclusion must be individually false with the antecedent implying the conclusion. However, in classical logic, any conditional with a false antecedent is trivially true. A more interesting question is whether a counterfactual is reasonable or *verifiable* i.e. does the conclusion follow from the antecedent. Many approaches to the verifying counterfactuals have been suggested. The majority follow the spirit of

Given: *if A then B*

i. set up a pretence space where A is true.

ii. remove any proposition which contradicts A.

iii. construct meaning of the counter factual by verifying that A entails B within the pretence space. (Brée, 1982)

Step ii is essential since if the antecedent is added to the set of true propositions then a contradiction will arise (since the negation of the antecedent is present). Clearly, it is not sufficient to remove just the negation of the antecedent: in addition, any proposition which entails the negation of the antecedent must also be removed. This step must be repeated till the set of true propositions is consistent. An obvious answer would be to remove any proposition which would be false if the antecedent was true. However, as Goodman (1947) points out, this process can be expressed as

another counterfactual and so a circularity is introduced. Goodman termed this the *contenability* problem i.e. the problem of specifying which propositions were contenable with the counterfactual.

Further evidence refuting the above is that counterfactuals often can not be verified in any classical sense: e.g. a counterfactual might describe an absurd or contradictory scene. For example, counterfactuals are often metaphorical:

(2) "If Clinton was the Titanic, the iceberg would sink[1]".

In our analysis of counterfactuals in the *CobuildDirect* corpus we found that figurative counterfactuals are common. Such a counterfactual cannot be evaluated without reasoning about the conceptual metaphors (Lakoff, 1993) employed by the counterfactual. Rather than treating such examples as a special case, in this paper we will argue that the same need for mapping between domains via established correspondences involved in reasoning about conceptual metaphors (Barnden & Lee, 1999) is also essential to reasoning about counterfactuals. To this end, we have implemented an approach to counterfactual reasoning within an AI system, ATT-Meta, originally designed for reasoning about metaphors and nested belief spaces.

Counterfactuals are often blends of several disparate domains, for example:

(3) "If Julius Caesar was in charge of the Korean War then he'd have used the atom bomb."
(Quine, quoted by Lewis, (1973))

In example 3, domain knowledge about the Roman Empire and the Korean War is blended in the same counterfactual space. The problem is specifying what knowledge is imported from which domain: the counterfactual appears quite reasonable if Caesar's ruthlessness is imported from the Roman Empire domain but not if his knowledge of military technology is also considered. Clearly any theory of counterfactuals must be able to specify which particular facts are imported from the domains to verify the counterfactual.

We claim that the contenability problem arises due to a mistaken assumption in classical semantics that any natural language utterance can yield a well-defined and context free set of propositions which capture the total meaning of the utterance and that these propositions can then be evaluated independently of the conceptual structure of the utterance or its context.

In the remainder of this paper, we will present an alternative account based within the tradition of cognitive linguistics but implemented within a computer program which stresses the role of metaphor, domain mapping, and uncertainty in reasoning about language. We support Fauconnier's claim (1997) that the structure of counterfactuals is not truth functional (i.e. a case of entailment from an alternative set of premises) but that rather counterfactuals are used to draw analogies from some alternative conceptual space. Given a counterfactual, our general approach involves setting up a separate reasoning space (henceforth termed the counterfactual space) where the antecedent is asserted as true. The counterfactual is then verified by the importation of facts from the domains referred to in the utterance plus whatever reasoning is necessary within the counterfactual space.

One aspect often overlooked in philosophical treatments of counterfactuals is the actual purpose a counterfactual plays within any wider context (i.e. its perlocutionary effects). We will present a goal-directed approach based on the drawing of analogies between the contents of the counterfactual space and the domain under discussion.

1. Turner and Fauconnier (in press) discuss this example in greater depth. However, they focus on the blending of the two domains rather than the statement's counterfactual structure.

2 An outline of the theory

A counterfactual expresses three things: A space builder, an antecedent and a conclusion. A space builder (Fauconnier, 1997) is a grammatical expression which either creates a new reasoning space or shifts attention to a previously connected space. In terms of counterfactuals, the pattern:

if *antecedent* then *conclusion*

instructs the creation of a counterfactual space. This space, though internally consistent, is not required to be consistent with any other logical space, including the set of true propositions.

Within the counterfactual space, the antecedent is assumed to be true. Verification proceeds by showing that the conclusion follows from the antecedent within this counterfactual space. This involves the selective importation of facts from the conceptual domains referred to by the statement (henceforth termed the *source domains*). In the case of literal counterfactuals, only one source domain is involved e.g. example 1 refers to John's investments in particular and the financial market in general. We could refer to this source domain as the "financial" domain and allow the importation of facts from this domain into the counterfactual space.

In the case of metaphorical or blended counterfactuals however, several domains might be involved. E.g. in example 3, the domains of President Clinton's impeachment trial and the voyage of the Titanic are both drawn upon and mapped over via conceptual metaphors. In Example 4., domain knowledge of both the Roman Empire and the Korean war is required.

Clearly, the importation of facts from any source domain runs the risk of inconsistency within the counterfactual space. We handle conflict resolution by assigning all propositions with a qualitative uncertainty and resolve conflicts via a measure of specificity. As we shall see, conflict resolution is a key part of our approach to counterfactual reasoning.

Once a counterfactual has been verified, its connotation outside the counterfactual space can be established. Such a connotation is due to a pragmatic inference extra to but based on the counterfactual reasoning already done in verification. Connotations are established by the drawing of an analogy between the counterfactual space and the current situation (henceforth, the target space). In Section 8 we will present a method for drawing such analogies.

3. ATT-Meta

ATT-Meta is a rule-based reasoning system. For convenience, we will refer to reasoning (sub)goals, provided facts and derived propositions all as *hypotheses*. Hypotheses are actually (Barnden et al. 1994) terms in a first-order, episode-based logic, but for simplicity of presentation they will be portrayed here couched in natural language. This paper will only detail the concepts relevant to the current work but further details can be found in (Barnden & Lee, 1999).

ATT-Meta is capable of both simulative reasoning about beliefs and metaphorical reasoning. Nested reasoning spaces are allowed to facilitate simulation of other agents. Two types of space are distinguished: simulation-pretence cocoons and metaphor-pretence cocoons. Simulation-pretence cocoons are used to model the beliefs of other agents and may be nested. Metaphor-pretence cocoons are a special type of simulation pretence cocoon where the agent modelled is a hypothetical agent who assumes that the manifested metaphor is literally true.

Metaphors in discourse are handled by reasoning about the consequences of treating the metaphor as literally true in the metaphor-pretence space and then mapping hypotheses from the metaphor-pretence space to the target space in which the metaphorical utterance occurs. Counterfactual spaces are a specialised type of metaphor-pretence space where the mapping between source and target domains are non-specialised importation rules. Such rules are

described in Section 5.

At any time any particular hypothesis is tagged with a qualitative certainty level, one of certain, presumed, suggested, possible or certainly-not. ATT-Meta applies its rules in a back chaining style.

Given a reasoning goal to prove, ATT-Meta attempts to supply a certainty level to it and to its complement. It does this by the use of rules to generate sub goals in the normal way. Goals can of course also be satisfied by provided facts. When a rule application successfully supports a goal, it supplies a level of certainty to it, calculated as the minimum of the rule's own certainty level and the levels picked up from the hypotheses satisfying the rule's antecedent. The rule's certainty level thus acts as an upper bound on the certainty that the rule can supply to its conclusion. The rule only supplies support to the goal in this way if all those certainty levels are at least suggested (because every hypothesis starts off as possible anyway). When several rules support a hypothesis, the maximum of their certainty contributions is taken.

In addition, the system contains a modified type of justification-based truth-maintenance mechanism for propagating levels of certainty around via rule-applications which is represented by a hypothesis graph which is essentially a dynamically changing AND/OR graph defined by the rule-applications that have been created. There are often cycles in the hypothesis graph. In such a case, the system iteratively settles to a consistent set of hypothesis certainty levels. This process is proven to always terminate.

4 Basic Conflict Resolution

When both a hypothesis and its complement are supported to level presumed, conflict-resolution will take place.

In this case, the system attempts to see whether one hypothesis has more specific evidence than the other. The former is then the "winner." If the system determines a winner, it downgrades the certainty level of the loser to suggested, keeping the winner's at presumed. If it cannot determine a winner, it downgrades both to suggested. It is therefore conservative in its conflict resolution, avoiding arbitrary decisions between hypotheses. For instance, in the Nixon diamond it will end up attaching level suggested to the hypotheses that Nixon is and is not a pacifist.

Specificity comparison is a commonly used heuristic for conflict-resolution in AI (see, e.g., Loui et al. 1993) although serious problems remain in coming up with adequate and practical heuristics. ATT-Meta's specificity comparison depends on what facts the two conflicting hypotheses rely on and on derivability relationships between the hypotheses supporting the conflicting hypotheses. The derivability aspect is analogous to derivability aspects of other schemes for specificity, and has as a special case the ancestor-closeness heuristic central to most semantic networks. The fact aspect relies on the intuition that if one of the conflicting hypotheses relies on more facts (i.e. user supplied propositions with certainty levels of presumed or certain) than the other then it has more specific evidence ("more" in a set-inclusion sense). More precise details of specificity-comparison are not important for understanding the rest of the paper but are specified in further detail in (Barnden et al. 1994).

5 Literal Counterfactual Reasoning

We will illustrate our approach to counterfactuals first with a relatively mundane example. Reconsidering example 1 and assuming that the following facts are true and represented as domain facts about John's financial dealings:

F1 John has shares to sell. (Certain)

F2 John has not sold his shares. (Certain)

F3 The value of his shares has increased since he bought them. (Certain)

F4 Selling shares results in not owning the shares. (Certain)

F5 Selling shares results in gaining the current value of the shares. (Certain)

We assume a general rule that things that are said are usually presumed to be true.

Locution: Any proposition uttered is true (presumed)

It is important to note that despite this rule, it is possible for an uttered hypothesis to be suppressed to a lower level of certainty or even falsified if there is sufficient conflicting evidence.

Given that the counterfactual "If John had sold his shares, he would have made a profit" is uttered, e.g.

U1: Uttered(if John had sold his shares, then John made a profit) (Certain)

then a counterfactual space must be built. This space is linked from the current topic of discussion (John's finance). Within this counterfactual space, the following proposition is asserted:

C1 John sold his shares (Certain)

The task, therefore, is to prove that the conclusion of the counterfactual, C2 follows from C1:

C2 John made a profit.

Since the counterfactual space only holds the asserted proposition C1, more information is required. This is achieved by an importation rule which allows facts to be imported into the counterfactual space from any connected source domain.

Importation: Import any fact from any connected source domain into the counterfactual space as required but at a lower level of certainty.

This allows for all of the facts in the financial domain to be imported but at a lesser degree of certainty: certain facts in the source domain are merely presumed in the counterfactual space; presumed facts in the source domain are merely suggested. This ensures that any contradictory fact to C1 is suppressed. It is important to note that the importation itself is an hypothesis with a presumed certainty rating.

In order to verify the counterfactual, ATT-Meta attempts to prove C2 follows from C1 in the counterfactual space. This involves establishing the certainty level of C2 and its complement in this space by the importation of hypotheses from the financial domain. To this end, the following facts are imported as counterfactual hypothesises:

C3: The value of his shares has increased since he bought them. (Presumed)

C4: Selling shares results in gaining the current value of the shares. (Presumed)

C5: Uttered(if John had sold his shares, then John made a profit) (Presumed)

This situation is shown in Figure 1. Following the importation of facts from the source domain, it is possible to show that the conclusion of the counterfactual presumably follows from the antecedent via two routes: first, via the locution rule, C5 plus C1 supports C2 being presumed; secondly, that C1 combined with C3 and C4 supports C2 being presumed. It is this second line of reasoning which actually verifies the counterfactual. It is important to note that both supports for C2 are important: clearly, a counterfactual cannot be verified merely by the fact that it has been said. Instead, verification relies on C1 supporting C2 independently of whether the counterfactual has been uttered. However, as we shall see, additional support is needed in the case of conflict resolution in more complex examples.

The above demonstrates the basic mechanism of counterfactual reasoning. In the next section, we show that the above approach can be extended to deal with blending phenomena.

6 Blending within Counterfactuals

To recap example 3:

(3) "If Julius Caesar was in command during the Korean war then he'd have used the atom bomb."

Unlike example 1, example 3 refers to concepts in two disparate domains: the Roman Empire and the Korean War. Therefore, verification must blend aspects of both domains in counterfactual space. However, what aspects of each domain are imported into the counterfactual space and how they are blended is problematic. Our approach is simple: we allow importation from either domain but use general conflict resolution in an attempt to suppress any hypothesis which is not consistent with the verification of the counterfactual. For example, assume the following hypotheses are true and represented as domain facts in a "Roman Empire" domain.

R1 Caesar was in command (Certain).
R2 Caesar was ruthless (Certain).
R3 A weapon used was the catapult (Certain).
R4 The catapult was the most devastating weapon (Certain).

Similar assume the following hypotheses are true and represented as facts in a "Korean war" domain.

K1 McArthur was in command of the American army (Certain).
K2 A weapon used was the Atom Bomb (Certain).
K3 The Atom Bomb was the most devastating weapon (Certain)
K4 McArthur did not use the Atom Bomb (Certain).

Finally, assume that there are general rules which apply to all domains:

G1 If a commander is ruthless, then he will use the most devastating weapon (Certain).
G2: Only one weapon can be the most devastating. (Certain)

Given the above facts, assume the counterfactual is uttered while discussing the Korean war, so that:

U1 Uttered(Caesar was in command during the Korean War, entails that Caesar used the atom bomb) (Certain)

is asserted within the "Korean war" domain. The verification of this counterfactual prompts the creation of a counterfactual space linked from the "Korean war" domain. Since the counterfactual

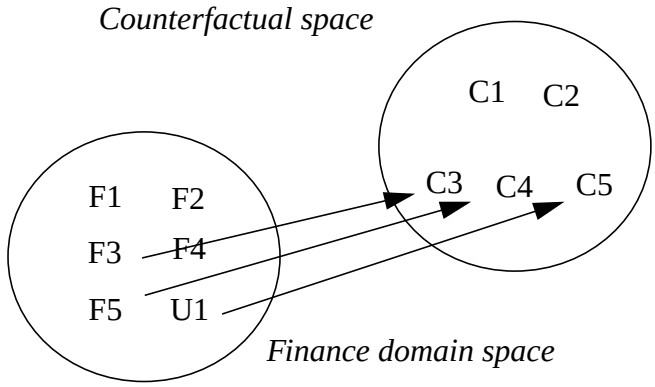

Figure 1: *"If John had sold his shares, he'd of made a profit"*.

also explicitly refers to concepts from the “Roman Empire” domain, this source domain is also linked. Within this counterfactual space, C1 is asserted:

C1 Caesar was in command of the American army (Certain)

In order to verify the counterfactual, the theorem prover must prove that the conclusions, C2, follows from C1:

C2 Julius Caesar used the atom bomb.

ATT-Meta attempts to prove C2 follows from C1 in the counterfactual space. This involves establishing the certainty level of C2 and its complement in this space by the importation of hypothesises from both source domains. To this end, R2 is imported from the “Roman Empire” domain, and K2, K3 and U1 are imported from the “Korean war” domain to form the following hypothesises in the counterfactual space:

C3 Caesar was ruthless. (Presumed)

C4 The Atom Bomb was a weapon used in the Korean War.
(Presumed)

C5 The Atom Bomb was the most devastating weapon. (Presumed)

C6 uttered(Caeser was in command of the Korean War, entails that Caesar used the atom bomb) (Presumed)

Two lines of reasoning support C2: First, C1, C3, C4 and C5 plus with the domain independent rule G1, show that C2 is presumed (and verify the counterfactual). In addition, C6 via the locution rule, plus C1 support C2 as presumed.

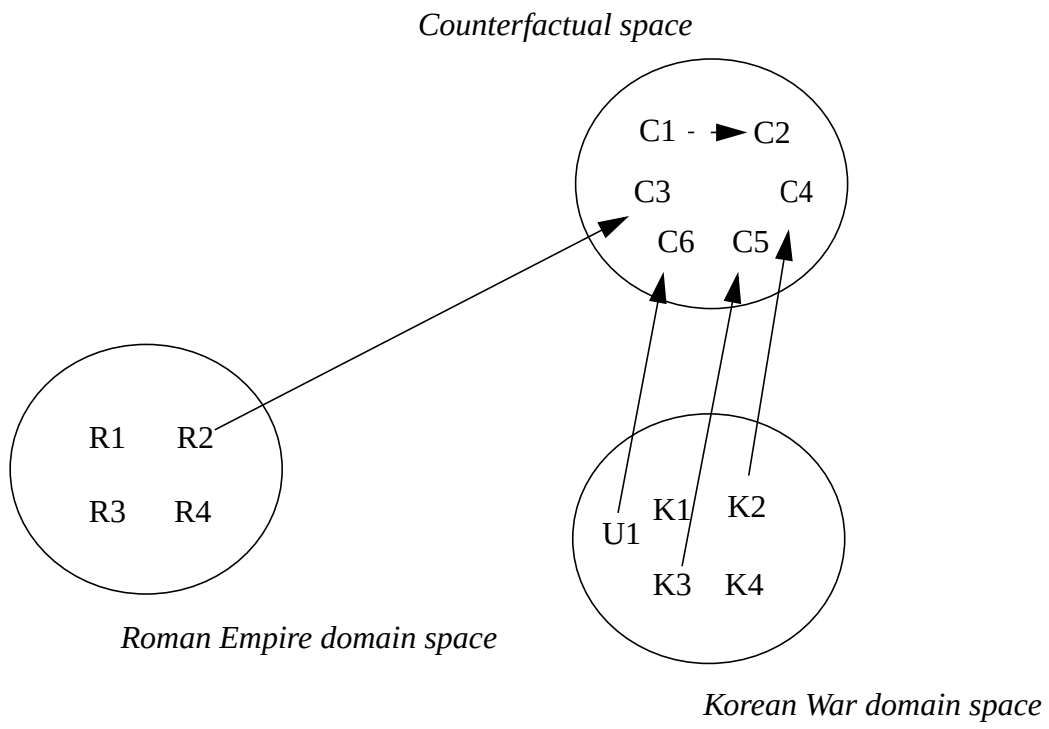

Figure 2. “If Julius Caesar was in command during the Korean war then he’d have used the atom bomb”.

7 Uncertainty in Counterfactual Reasoning

It is worth considering what verification means. Clearly, verification of a counterfactual is weaker than logical validation. Instead, verification is a judgement of reasonability involving the subjective importation of facts. This judgement can be disputed. As Quine points out, a perfectly reasonable reply to example 3 would be:

(4) "...no, he'd have used the catapult"

Such an utterance denies the verification of the previous counterfactual and makes any counterfactual comparison between the Roman Empire and the Korean war meaningless. Clearly, any theory of how counterfactuals work must account for this. Our analysis is as follows: the reply in 4. expresses a new counterfactual which provides the original antecedent with a new conclusion, i.e.

(5) "If Julius Caesar was in command during the Korean war then he'd have used the catapult."

This counterfactual must be evaluated independently of the original counterfactual - despite both sharing the same antecedent being linked to the same source domains, each space appeals different information.

U2 Uttered(Caeser was in command of the Korean War entails Caesar used the catapult) (Certain)

This utterance creates a separate counterfactual space where C1 is asserted and the new conclusion C7, is verified.

C7 Caesar used the catapult.

This involves the importation of R2, R3, R4 from the "Roman Empire" domain, and U2 from the "Korean war" domain, forming the following hypotheses in the second counterfactual space:

C8 Caesar was ruthless (Presumed).

C9 A weapon used was the catapult (Presumed)

C10 The catapult was the most devastating weapon (Presumed)

C11 uttered(Caeser was in command of the Korean War entails Caesar used the catapult) (Presumed)

As before, the resulting contents of the counterfactual space are sufficient to verify the counterfactual expressed in example 6. The resulting counterfactual spaces are shown in Figure 2.

Each counterfactual space need be only internally consistent. Moreover, the set of hypotheses imported in each space block the importation of any hypothesis which contradicts the verification. For example, before U2 is uttered, any hypothesis importing R4 into the original counterfactual space will be suppressed due to there being more specific evidence for the counterfactual (that is, the facts necessary to verify the counterfactual plus the utterance of that counterfactual). However, once U2 is uttered, the two competing counterfactuals have an equal amount of specific evidence. In which case, ATT-Meta's conservative conflict resolution would downgrade all conflicting evidence in both spaces to merely suggested.

8 Connotations of Counterfactuals

A counterfactual describes an alternative world where, though false in the real world, its antecedent and conclusion are true. Clearly, for a counterfactual to have any relevance in context, it must have some connotation which can be transferred to the current discourse (the target domain).

As is the case with metaphors, for there to be a connotation in the target space, there must be a mapping from the counterfactual space to the target space. However, Unlike the case of meta-

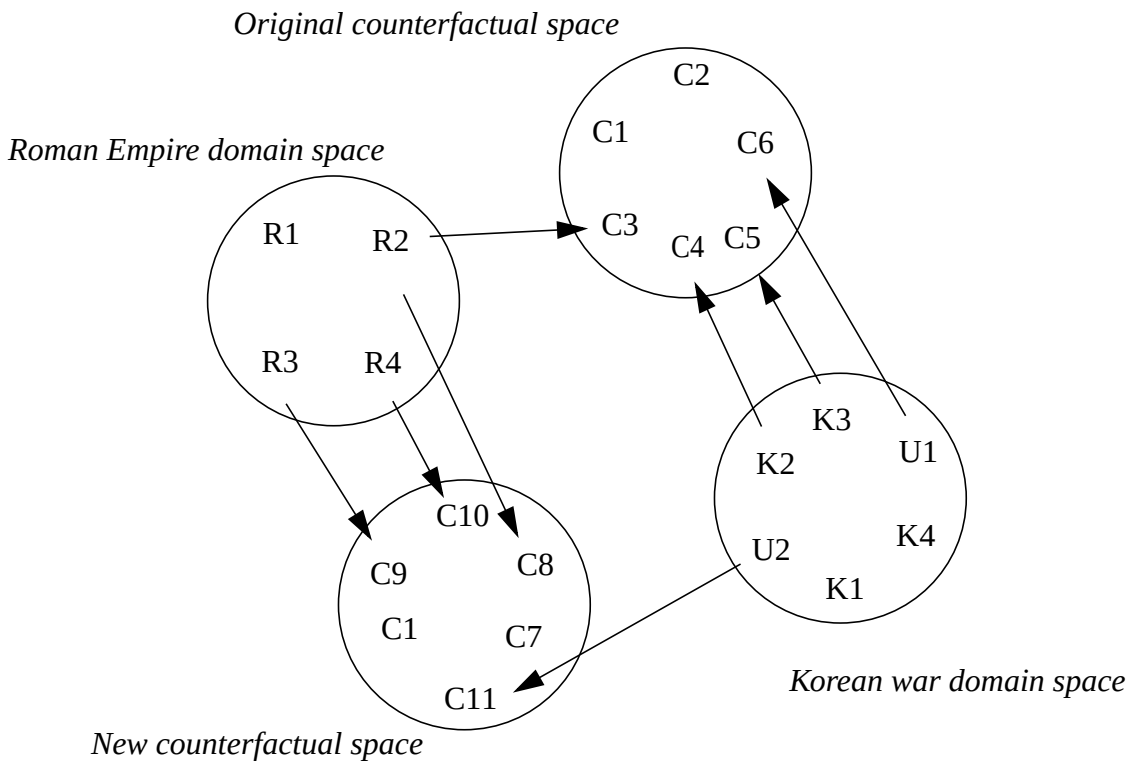

Figure 3: *Comparison of two alternative counterfactuals*

phors, where mappings between the source and target domains are familiar to the hearer as conceptual metaphors, counterfactuals present novel situations and therefore must rely on domain-independent analogical mappings between the counterfactual space and its target domain. We do not claim to have a complete set of rules for analogical mapping and this is a topic for further research. However, the following rule is sufficient for the examples discussed above.

Analogous Inference Mapping: An inference chain in counterfactual space can be mapped to an analogous inference chain in the target domain.

The term analogous inference chain needs to be clarified. Clearly, the same inference cannot be made in both spaces since the sets of hypotheses in both spaces are different and contradictory. Instead, an analogous inference is a transformation where analogous entities are substituted and the inference chain transposed.

Analogous hypotheses share the same predicate structure, for example, C1 and K1 are analogous (in that they both identify the commander during the Korean war in their respective spaces). The above can be illustrated using example 3. Given that the counterfactual "if Caesar had been in command during the Korean War, he'd have used the atom bomb" has been verified as detailed in Section 7 and the counterfactual space contains the inference chain:

C3 Caesar was ruthless. (Presumed)

supports

C2 Caesar used the atom bomb (Presumed)

This inference chain can be mapped back to the target space of the Korean war. Given that McArthur (the Commander of the American War effort in the target space) did not use the atom

bomb, the only applicable transformation is to assume the contra-positive of the above inference chain, i.e.

K4 McArthur did not use the Atom Bomb (Certain).

supports

K5 McArthur was not ruthless (Presumed)

Therefore, the connotation of the counterfactual is the inference of K5. Clearly the connotation of any counterfactual must be sensitive to the context in which the utterance expressing the counterfactual is made. This is captured in our approach in that any analogy inspired transformation is dependent on the set of hypotheses in the target space. For example, supposing that the counterfactual was uttered during a discussion of the Roman empire then the only possible connotation would stress how ruthless Caesar was.

9 Conclusions

We have presented an account of counterfactual reasoning which stresses the roles analogy and the mapping of information between conceptual domains play in understanding. The account has two parts. The first, verification, involves the subjective importation of information into a conceptual reasoning space plus conflict resolution to determine the set of hypotheses required to justify the counterfactual. The second, determining the connotation of the counterfactual involves the drawing of analogies between the counterfactual's reasoning space and the current discourse domain.

References

Barnden, J.A., Helmreich, S., Iverson, E., & Stein, G.C.(1994). An integrated implementation of simulative, uncertain and metaphorical reasoning about mental states. In J. Doyle, E. Sandewall, P. Torasso (Eds), Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference,} pp.27--38. San Mateo, CA: Morgan Kaufmann.

Barnden, J.A. & Lee, M.G. (1999). An implemented context system that combines belief reasoning, metaphor-based reasoning and uncertainty handling. *Procs. Second International and Interdisciplinary Conference on Modelling and Using Context* (Context'99), Trento.

Brée, D.S. (1982) Counterfactuals and Causality. *Journal of Semantics* 1.2.

Fauconnier, G. (1997). *Mappings in Thought and Language*. Cambridge University Press.

Goodman, (1947) The problem of counterfactual conditional sentences. *Journal of Philosophy*, 44.

Lakoff, G. (1993) The Contemporary Theory of Metaphor. In A. Ortony (ed.) *Metaphor and Thought*, 2d ed., 202-51. Cambridge University Press.

Lewis, D.K. (1973) *Counterfactuals*. Cambridge, Mass., Harvard University Press.

Loui, R.P., Norman, J., Olson, J. & Merrill, A. (1993). A design for reasoning with policies, precedents, and rationales. In *Proc. of the Fourth International Conference on Artificial Intelligence and Law*, pp.202--211. New York: ACM.

Turner, M. & Fauconnier, G. (in press) Metaphor, Metonymy, and Binding. In A. Barcelona (eds.) *Metonymy and Metaphor*. Mouton de Gruyter.