Mirlog: A logic for multimedia information retrieval

Carlo Meghini, Fabrizio Sebastiani and Umberto Straccia
Consiglio Nazionale delle Ricerche
Istituto di Elaborazione dell'Informazione
Via S. Maria 46 - 56126 Pisa, Italy

E-mail: {meghini,fabrizio,straccia}@iei.pi.cnr.it

Abstract

This chapter presents a logic for the retrieval of multimedia information, whose ultimate goal is to model retrieval as an uncertain logical inference, in accordance to the logic-based view of retrieval. The logic being presented is the product of a number of extensions to a Description Logic, which constitutes the kernel of our model. Each extension is meant to capture an important aspect of the retrieval endeavour that is not adequately dealt with by the kernel logic. The extensions are: a 4-valued semantics, aiming at capturing relevance in retrieval; closure assertions, aiming at allowing closed-world reading of selectively specified parts of a document base; fuzzy assertions, aiming at handling the uncertainty intrinsic in retrieval. The resulting logic is to be understood as a modelling retrieval tool, which can be used for the specification and the rapid prototyping of applications.

Contents

1 Introduction

The development of retrieval models has been a major concern of the information retrieval community for the last two decades. As a result of this effort, we now have a few well established and widely known models, around which famous information retrieval systems have been built and put at work on real applications. These models are based on different views of the retrieval process, but they all share two common features: first, they have been developed for and mostly applied only to textual documents; second, they adopt an indirect approach, based on statistical properties of keywords, to the central problem of information retrieval: capturing document contents. Both these features were dictated by the context in which the relevant research took place. As for the former, text was the only medium that could be automatically processed in an efficient way until a few years ago. As for the latter, the choice of a "surface" approach to capturing meaning was imposed by three factors:

1. the sheer size of major applications, where collections of thousands or millions of textual objects were addressed, thus making automatic extraction of document representations a necessity;

2. the lack of tools for automatically extracting more faithful renditions of document semantics;

3. the lack of theories that give a satisfactory explanation of what document semantics really is.

Concerning points 2 and 3, the possibility of automatically extracting the meaning of a text by simply extracting the semantics of each sentence and combining the results is both conceptually naïve and practically unattainable. Although the formal semantics of significant fragments of natural language is now well understood and derivable in an automatic way, there are still parts of discourse that resist to automatic treatment. And even granted that a program could come up with the semantic structure of any piece of text, the resulting logical theory would raise computational problems of unattackable complexity (see e.g. [27]).

Things go from bad to worse once one moves from the case of textual documents, to that of documents pertaining to other media. The scenario in which IR systems are supposed to operate has undergone major changes in the recent past, as media other than text have appeared on the scene, giving rise to applications with features and requirements radically different from those of traditional text-based systems, and therefore calling for different methods of tackling the problems involved. In this case, the possibility of automatically extracting *any* sort of meaning from non-textual documents seems to be at present beyond the ability of any computer program. The practical consequence of this is that traditional models based on surface meaning are not immediately applicable to multimedia retrieval, unless one is willing to abandon the idea of performing retrieval by semantic content.

1.1 The approximated content paradigm

We believe that when we are confronted with the problem of grounding multimedia IR in a theory of meaning, there does exist an alternative between the "surface" kind of meaning based on keyword statistics, and the "deep" kind of meaning that is still in the realm of the ineffable. In particular we believe that for the purpose of retrieval the "deep" meaning of a document could be reasonably *approximated* by means of expressions of a formal language that, while accounting for the intensionality of semantics as opposed to the extensionality inherent in a statistical approach, escapes the elusive character of "deep" semantics. This approach is evoked by a now classic textbook in IR [52, page 189]:

> It has never been assumed that a retrieval system should attempt to "understand" the content of a document. Most IR systems at the moment merely aim at a bibliographic

> search. Documents are deemed to be relevant on the basis of a superficial description. I do not suggest that it is going to be a simple matter to program a computer to understand documents. What is suggested is that some attempt should be made to construct something like a naïve model, using more that just keywords, of the content of each document in the system. The more sophisticated question-answering systems do something very similar. They have a model of the universe of discourse and can answer questions about it, and can incorporate new facts and rules as they become available.

Approximate meaning nowadays still needs to be conveyed by a human indexer, who specifies it according to his understanding of the document contents in the context of a domain of discourse, but research in natural language understanding and knowledge representation is now mature for producing robust tools that might automate this task to a large extent. In particular, the approximate meaning scenario, which is commonplace in libraries as we know them from everyday experience, requires two fundamental tools to be realized: first, a language that the indexer can use for expressing the approximate meaning of a document; second, a retrieval engine able to exploit such meaning in the retrieval process. There has been considerable progress in recent years concerning tools of this kind. Of special interest to the present paper, research in the knowledge representation area has yielded Description Logics (DLs), term-oriented logics whose expressive power and inferential abilities are, as we will argue later, to a large extent adequate to the approximate meaning approach that we have sketched. A wide range of results are available on the computational aspects of these logics, and this allows both to classify the decision problems of these logics from the complexity point of view and, on a more practical side, to confidently develop automated reasoning systems based on them.

The availability of these tools puts us in a position of moving significant steps in the direction pointed to by [52]; as a matter of fact, the above quoted paragraph can be seen as the *manifesto* of our approach. We aim at formulating an IR model where the meaning of documents, although in an approximated form, is explicitly represented by means of sentences of a DL, thus marking a substantial difference from keyword-based models. Moreover, the terms occurring in document representations can additionally be described and interrelated by means of assertions of the same logic, having the syntactical status of definitions and playing the role of a lexicon. Finally, domain knowledge can be expressed also via assertions of the logic. All these kinds of knowledge, and possibly others, such as document profiles, are brought to bear on document retrieval by modelling the latter in terms of the entailment relation of the logic. The resulting model goes very close to the question-answering sort of system mentioned in the above passage; more precisely, we can think of it as a system capable of answering questions regarding what documents are about, thereby gaining the ability to do semantic content-based document retrieval. How effective the system is will strictly depend on the quality of the approximation of meaning, the richness of the lexicon and the completeness of domain knowledge.

1.2 Building blocks

The logic that we have designed for multimedia document retrieval can be seen as consisting of a core, relatively unsophisticated logic, to which features are added incrementally in order to make it more respondent to the complexity of the IR task.

In Section 2 we introduce the core logic, which may be seen as representing a first solution to the problem of addressing the conditional reasoning part of van Rijsbergen's proposal. This tool is the description logic $\mathcal{ALC}$, a logic that, while essentially based on the semantics of first order logic (FOL), has a significantly different language than that of FOL. This language is oriented to the representation of classes of structured objects, allowing to view the problem of retrieval as one of deciding whether an object (a document) is an instance of a class (a query). As we show in Section 2, and as we more fully argue in [36], this allows a more natural account of retrieval, and also allows both documents and queries to be given rich, structured representations that describe them under multiple viewpoints. The same "object-oriented" language may be used to give structured representations of the meaning of the words that occur in document and query representations, i.e. to formally represent dictionary or thesaural entries. As a consequence, words are no longer seen as uninterpreted tokens exclusively characterized by their occurrence ratio in documents, but as *intensional* objects, i.e. objects having a meaning that systematically affects the meaning of the sentences in which they occur, and that by virtue of their meaning convey information. We feel that this collection-independent meaning has to be explicitly represented and used in order to make full sense of documents and queries, and allow thereby effective retrieval. Our view of logic-based IR can thus be seen as a move from an extensional view of meaning to an intensional one.

In Section 3 we go one step further in addressing the conditional reasoning issue, and propose a semantics that better mirrors the classic IR notion of relevance. The issue we tackle in particular is that of accepting as indicative of relevance only those implications whose premise contains information relevant to the conclusion. This condition is identified as the requirement that evidence supporting the conclusion be explicitly present in the premise; this is a stronger requirement than that of also being content with the absence of negative evidence, a weaker requirement that may be seen as informing the approach initially taken in Section 2. This switch of focus is accomplished by abandoning classical logic in favour of *relevance logic*, which in turn implies abandoning classical two-valued semantics in favour of 4-valued semantics. The impact of this modified logic on IR is also thoroughly discussed in [50].

In Section 4 we tackle the long-standing issue of *closed- vs. open-world reasoning* in IR, i.e. the problem of deciding whether in reasoning about IR all that is *known* about a given document, or word, or topic, should be equated with all that is *true* about it. If this is the case, the logic should be modified so as to make it behave according to the closed world assumption. We argue that there are different types of knowledge whose representation contributes to deciding retrieval, and that, while some of them are naturally interpreted in a closed way, some others require open-world reasoning (see also [38] for a fuller discussion of this point). In order to account for this phenomenon, we extend the logic with capabilities for *selective closed-world reasoning*; this means the possibility of reasoning in closed-world style with some items of knowledge, and in open-world style with others. A sophisticated, fine-grained mechanism of *meta-assertions* is provided that allows to indicate that a given individual or a given predicate symbol requires a closed-world reading.

Section 5 finally tackles the problem of adding *reasoning about uncertainty* to the framework for conditional reasoning developed so far. What we need is a framework in which, rather than deciding *tout court* whether a document is relevant to a query, we are able to *rank* documents according to how strongly the system believes in their relevance to queries. Section 5 discusses an extension of the logic presented so far to the case of fuzzy reasoning .

It is important to remark that all of the developments listed so far are not studied merely from the point of view of the logical language and semantics adopted. In fact, the logic presented in this paper is endowed with a calculus for reasoning in it (that we have developed incrementally in the form of a sequence of increasingly more powerful calculi, to parallel the development of the logic), for which we have obtained soundness, completeness and complexity results. Because of space limitations, we will not go into the details of our proof system, but will confine ourselves to an overview, in Section 6. Finally, Section 7 concludes.

2 Syntax and classical semantics

Following [36], the kernel of our model is based on a logic from the class of Description Logics[1] (DLs), the descendants of the frame-based knowledge representation languages of the late seventies (see e.g. [5]). The basic entities of the language of a DL are: *individuals* (denoted by the letter a, with optional subscript), representing objects of the application domain; *concepts* (letter C, with optional subscript), representing classes of objects; and *roles* (letter R, with optional subscript), representing binary relations between objects. In the same way as in predicate logic complex formulae are built out of predicate symbols via connectives and quantifiers, in DLs complex concepts and roles are built out of unary predicate symbols (aka *primitive concepts*), binary predicate symbols (*primitive roles*) and individuals via *concept-* and *role-forming operators*, respectively. For example, the complex concept

$$\texttt{Paper} \sqcap \forall \texttt{Author}.\texttt{Italian}$$

is obtained by combining the primitive concepts `Paper` and `Italian` and the primitive role `Author` by means of the conjunction ($\sqcap$) and the universal quantification ($\forall$) operators; under the intended interpretation of these symbols, the concept denotes the set of papers whose authors are all Italians[2]. Concepts and roles are collectively called *terms*. From the syntactical point of view, Mirlog is the logic $\mathcal{ALC}$ extended with primitive role negation. $\mathcal{ALC}$ is a significant representative of the best-known and most important family of DLs, the $\mathcal{AL}$ family[3]. Concepts and roles[4] of Mirlog are formed out of primitive concepts (letter A) and primitive roles (letter P), according to the following syntax rules:

$$\begin{array}{rrl}
C \quad \longrightarrow & \top \mid & \text{(top concept)} \\
 & \bot \mid & \text{(bottom concept)} \\
 & A \mid & \text{(primitive concept)} \\
 & C_1 \sqcap C_2 \mid & \text{(concept conjunction)} \\
 & C_1 \sqcup C_2 \mid & \text{(concept disjunction)} \\
 & \neg C \mid & \text{(concept negation)} \\
 & \forall R.C \mid & \text{(universal quantification)} \\
 & \exists R.C & \text{(existential quantification)}
\end{array}$$

$$\begin{array}{rrl}
R \quad \longrightarrow & P \mid & \text{(primitive role)} \\
 & \neg P \mid & \text{(primitive role negation)}
\end{array}$$

As customary, we will use parentheses around concepts when the need for disambiguation arises.

[1] In order to follow the naming conventions adopted in the Description Logic community, the name of our logic should be something like "$\mathcal{ALC}^c_4$". However, partly because of the awkwardness of the candidate official name, partly to emphasize the context in which the logic has been developed (MIR here stands for Multimedia Information Retrieval), we prefer to call the logic "Mirlog".

[2] In this paper we follow the now standard "FOL-like" syntax of DLs, e.g. writing $\texttt{Paper} \sqcap \forall \texttt{Author}.\texttt{Italian}$ in place of the expression (**and** `Paper` (**forall** `Author Italian`)) encoded in the "Lisp-like" syntax of [36]. Also, we use the now standard name "*Description* logics" instead of the name "*Terminological* logics" used in [36].

[3] The DL we had employed in [36, 48], called Mirtl, embodied a different choice of operators than $\mathcal{ALC}$. The reason why we have now opted for a slight extension of $\mathcal{ALC}$ is twofold:

- in a recent study [9] we have found that Mirtl has, unlike $\mathcal{ALC}$ [47], bad computational properties;
- $\mathcal{ALC}$ is universally considered the "standard" description logic (as much as $\mathcal{K}$ is considered the "standard" modal logic) and is therefore regarded as the most convenient testbed for carrying out logical extensions and, in general, logical work of an experimental nature. Reverting to one's DL of choice may be taken as the very last (and usually straightforward) step in the development of a logical DL-based model.

[4] $\mathcal{ALC}$ only offers primitive roles.

Description logics have a clean, model-theoretic semantics, based on the notions of *truth* and *interpretation.* An interpretation $\mathcal{I}$ consists of a non empty set $\Delta^{\mathcal{I}}$ (called the *domain*) of *objects* (letter o) and of an *interpretation function* $\cdot^{\mathcal{I}}$ mapping primitive concepts into functions from $\Delta^{\mathcal{I}}$ to the set of classical truth values $\{t, f\}$[5] and primitive roles into functions from $\Delta_{\mathcal{I}} \times \Delta_{\mathcal{I}}$ to $\{t, f\}$. In compliance with the style of model-theoretic semantics, the interpretation of complex concepts and roles is obtained by appropriately combining the interpretations of their components. The (2-valued) semantics of Mirlog is the following:

$$
\begin{array}{rcl}
\top^{\mathcal{I}}(o) = t & & \text{for all } o \\
\bot^{\mathcal{I}}\ (o) = t & & \text{for no } o \\
(C_1 \sqcap C_2)^{\mathcal{I}}(o) = t & \text{iff} & C_1^{\mathcal{I}}(o) = t \text{ and } C_2^{\mathcal{I}}(o) = t \\
(C_1 \sqcup C_2)^{\mathcal{I}}(o) = t & \text{iff} & C_1^{\mathcal{I}}(o) = t \text{ or } C_2^{\mathcal{I}}(o) = t \\
(\neg C)^{\mathcal{I}}(o) = t & \text{iff} & C^{\mathcal{I}}(o) = f \\
(\forall R.C)^{\mathcal{I}}(o) = t & \text{iff} & \text{for all } o' \in \Delta_{\mathcal{I}},\ \text{if } R^{\mathcal{I}}(o, o') = t \text{ then } C^{\mathcal{I}}(o') = t \\
(\exists R.C)^{\mathcal{I}}(o) = t & \text{iff} & \text{for some } o' \in \Delta_{\mathcal{I}}, R^{\mathcal{I}}(o, o') = t \text{ and } C^{\mathcal{I}}(o') = t \\
(\neg P)^{\mathcal{I}}(o, o') = t & \text{iff} & P^{\mathcal{I}}(o, o') = f
\end{array}
$$

The interpretation of the concept $\mathtt{Paper} \sqcap \forall\mathtt{Author.Italian}$ is therefore such that: $(\mathtt{Paper} \sqcap\ \forall\mathtt{Author.Italian})^{\mathcal{I}}(o) = t$ iff $\mathtt{Paper}^{\mathcal{I}}(o) = t$ and for all $o' \in \Delta_{\mathcal{I}}$, if $\mathtt{Author}^{\mathcal{I}}(o, o') = t$ then $\mathtt{Italian}^{\mathcal{I}}(o') = t$, which corresponds to the informal reading suggested above.

Two concepts C_1 and C_2 are said to be *equivalent* (written $C_1 \equiv C_2$) when $t = C_1^{\mathcal{I}}(o)$ iff $t = C_2^{\mathcal{I}}(o)$ for all $o \in \Delta_{\mathcal{I}}$ and for all interpretations $\mathcal{I}$. This definition allows us to point to some *duality* in our set of operators. We may notice, for instance, that $\top$ and $\bot$ are dual, i.e. $\top \equiv \neg\ \bot$; similarly, $\sqcap$ is the dual of $\sqcup$, as $(C_1 \sqcap C_2) \equiv (\neg C_1 \sqcup \neg C_2)$, and $\forall$ is the dual of $\exists$, as $(\forall R.C) \equiv (\neg \exists R.\neg C)$.

The language of a DL also includes *assertions*, expressions relating concepts and roles to each other or to *individuals, i.e.* names of objects in the domain of discourse. Assertions of the former kind are called *definitions* and allow to state the existence of a specialisation ("more specific than") relation between concepts or between roles; for instance, the definition:

$$\mathtt{VisualDocument} \sqsubseteq \mathtt{Document} \sqcap (\exists\mathtt{Component.}(\mathtt{Image} \sqcup \mathtt{Video})) \tag{1}$$

asserts that visual documents are documents and have at least a component which is either an image or a video. Assertions of the latter kind are called *term* assertions and each of them may be a *concept* or a *role* assertion depending on the involved kind of term. Formally, given an alphabet $\mathcal{O}$ of symbols called individuals,

- a concept assertion is an expression of the form $C[a]$ where C is a concept and a is an individual; and
- a role assertion is an expression of the form $R[a, b]$ where R is a role and a, b are individuals.

For instance,

$$(\mathtt{Paper} \sqcap\ \forall\mathtt{Author.Italian})[\mathtt{o12}]$$

is a concept assertion stating that document named `o12` is a paper whose authors are all Italians, while the role assertion

[5] We adopt this notation in place of the equivalent but more widely used one that has the interpretation of a concept as a subset of the domain $\Delta^{\mathcal{I}}$, because it prepares the ground for the 4-valued semantics of Mirlog, to be introduced in the next section.

$$\texttt{Publisher[o12,Springer]}$$

states that the publisher of `o12` is `Springer`.

Assertional formulae[6] (letter γ) are Boolean combinations of term assertions realized according to the following syntactic rule (α denotes a term assertion):

$$\begin{array}{lrl}
\gamma \quad \longrightarrow & \alpha \mid & \text{(term assertion)} \\
& \gamma_1 \wedge \gamma_2 \mid & \text{(assertional conjunction)} \\
& \gamma_1 \vee \gamma_2 \mid & \text{(assertional disjunction)} \\
& \sim \gamma & \text{(assertional negation)}
\end{array}$$

The semantics of assertions is specified by extending the interpretation function $\cdot^{\mathcal{I}}$ to be an injection from $\mathcal{O}$ to $\Delta^{\mathcal{I}}$, according to the *unique name assumption*. In addition, the definition $C_1 \sqsubseteq C_2$ (resp. $R_1 \sqsubseteq R_2$) is *satisfied* by an interpretation $\mathcal{I}$ iff ${C_1}^{\mathcal{I}}(o)$ implies ${C_2}^{\mathcal{I}}(o)$ for all $o \in \Delta_{\mathcal{I}}$ (resp. ${R_1}^{\mathcal{I}}(o, o')$ implies ${R_2}^{\mathcal{I}}(o, o')$ for all $o, o' \in \Delta_{\mathcal{I}}$). Analogously, $\mathcal{I}$ satisfies $C[a]$ (resp. $R[a_1, a_2]$) iff $C^{\mathcal{I}}(a^{\mathcal{I}}) = t$ (resp. $R^{\mathcal{I}}({a_1}^{\mathcal{I}}, {a_2}^{\mathcal{I}}) = t$); moreover:

1. $\mathcal{I}$ *satisfies* an assertional formula $\gamma_1 \wedge \gamma_2$ iff it satisfies both γ_1 and γ_2;

2. $\mathcal{I}$ *satisfies* an assertional formula $\gamma_1 \vee \gamma_2$ iff it satisfies either γ_1 or γ_2, or both;

3. $\mathcal{I}$ *satisfies* an assertional formula $\sim \gamma$ iff it does not satisfy γ.

A set Σ of assertional formulae and definitions will be called a *knowledge base* (KB). A KB Σ *entails* an assertion $C[a]$ (written $\Sigma \models C[a]$) iff every interpretation satisfying all the expressions in Σ also satisfies $C[a]$. In this case, we will also say that $C[a]$ is a *logical consequence* of Σ. In what follows, we will sometimes be interested in discussing the case in which, given a KB Σ and two concepts C_1 and C_2, whenever Σ entails $C_1[a]$ it also entails $C_2[a]$ for all individuals a. In this case, we say that C_1 *is subsumed by* C_2 *in* Σ, and we write $C_1 \preceq_2^{\Sigma} C_2$. For example, the concept `VisualDocument` is subsumed by the concept `Document` in any KB containing (1). If C_1 is subsumed by C_2 in an empty KB, we simply say that C_1 *is subsumed by* C_2, and we write $C_1 \preceq C_2$. For example, the concept $\texttt{Document}\sqcap (\exists\texttt{Component}.(\texttt{Image} \sqcup \texttt{Video})$ is subsumed by the concept `Document`.

In [36] we have described in detail a methodology for giving representations of documents, queries and thesaural entries in terms of a description logic. In particular:

1. a document is to be represented as an individual; this individual will be the subject of a number of assertions; the concepts and roles of which the individual is asserted to be an instance will then altogether constitute the description of the document;

2. a query is to be represented as a concept; the intuitive meaning of this choice is that all documents represented by individuals that are recognised to be instances of this concept should be retrieved;

3. a thesaural entry is to be represented by means of a definition; the intended consequence of this choice is that the definition of a term be brought to bear whenever a document in whose representation a defined term occurs is considered, or whenever a query in whose representation a defined term occurs is issued.

Within these representations, documents may be considered from multiple viewpoints: the representation of a document may address its internal structure, its physical appearance, its semantic content

[6] Assertional formulae are not offered by DLs, including, of course, $\mathcal{ALC}$.

and its "profile" (i.e. the set of the identifying features of the document, such as authorship, date of production, etc.).

The information retrieval process may thus be viewed as deciding whether, given a KB containing document representations and thesaural entries, a concept C representing a query, and an individual a uniquely identifying a document, Σ entails $C[a]$.

3 A relevance semantics

Information retrieval is often characterized in terms of *relevance*: given a set of documents and a query, the task of IR is to retrieve those documents, and only those, whose information content is relevant to the information content of the query (aka user information need). The centrality of relevance and its elusive character, given its reliance on the unfathomable information content of documents and queries, is the main reason why the logical formalisation of IR is a non-trivial problem; what is relevant, that is, is decided by the user from session to session and from time to time, and is then heavily dependent on judgments where highly subjective and scarcely reproducible factors are brought to bear [3, 46]. The very possibility of a logical theory of IR is then dependent on the possibility of giving a *formal* definition of relevance capable of approximating the *operational* definition of relevance given above. In order to do so, it is of fundamental importance to at least identify part or all of those subjective and contingent factors that contribute to relevance, and wire them into one's adopted logic. Furthermore, we think that the addition of uncertainty on top of a calculus for conditional reasoning can indeed work as a "correction factor" for bridging the gap between the rigidity of logical calculi and the flexible, human-centered notion of relevance, as in principle it allows to fine-tune the system estimation of relevance as a function of contextual factors, user preferences and so on. We also think, however, that in order to arrive at a successful logical model of IR *every effort should be made in order to wire as much relevance as possible into the implication connective*, i.e. to design a calculus for (non-probabilistic) conditional reasoning where the factors that influence relevance, as perceived by the user, are taken into account. It is this consideration that motivates the research reported in this section.

3.1 Relation to other work

The history of logic has seen a flurry of logics motivated by the need to give a natural account of the implication connective. Quite interestingly, the accounts proposed by classical, modal and other logics, have been criticised on the account that they license, as theorems of the pure calculus, sentences that suffer from *fallacies of relevance*. In other words, some conditional sentences are theorems of the given logic *even if their premise is not relevant to their conclusion.* For instance, the sentence $(\alpha \rightarrow (\beta \rightarrow \alpha))$ (asserting that a true proposition is implied by any proposition) is a theorem of classical logic. And this should strike one as peculiar, in that the fact that β holds does not have any "relevance" to the fact that α holds! Among the first to put forth such a criticism, Nelson [39] argued that, in order for any conditional notion "$\rightarrow$" to be adequate, a sentence such as $\alpha \rightarrow \beta$ should be valid only if there is "some connection of meaning between α and β" (and this consideration should strike the IR theorist as familiar ...). To the surprise of many orthodox logicians who considered these issues to more properly belong to rhetoric rather than logic, the idea of a "connection of meaning between α and β" (or, more generally, the idea of α being *relevant* to β) has been shown to be amenable to formal treatment by the logicians who defined *relevance* (or *relevant*) *logics* [2, 19]. Relevance logics attempt to formalise a conditional notion in which relevance is a primary concern. By doing this, they challenge classical logic and its extensions in a number of ways, i.e. by introducing a new, non-truth-functional connective (denoted by "$\rightarrow$") into the syntactic apparatus of classical logic, by rejecting some classical rules of inference for classical connectives, and by changing the notion of validity itself by "wiring" into it considerations of relevance.

We think that, although they might not be a panacea for all the problems concerning the logical formalisation of IR, the insights provided by relevance logics are valuable to information retrieval. In fact, even a brief analysis of the motivations put forth by relevance logicians and by IR theorists, respectively, indicates a surprising coincidence of underlying tenets and purposes (see e.g. [24, Chapter 10]), much beyond the case of omonimy. Therefore, it seems just natural to think that, if we view retrieval as essentially consisting of a disguised form of logical inference [55], relevance logic and IR might constitute the theoretical side and the applied side of the same coin. This eventually calls for the *adoption of a relevance logic as the non-probabilistic kernel of a full-blown logic for IR.* Given that the description logics we have advocated in Section 2 are essentially based on classical logic, we intend to propose the switch to a *relevance description logic*; this will be the subject of the remaining part of this section.

As with modal logics, there are many relevance logics, each formalising a different notion of relevance. The relevance logic that we think best complies with the requirements of the IR world is the logic $\mathbf{E_{fde}}$, also called the *logic of first degree (tautological) entailments* [18]. This consists of the fragment of the famous relevance logics **E** and **R** that deals with *first degree entailments* only, i.e. pairs of propositional (classical) formulae separated by one "→" symbol. This logic seems well suited to formalise a state of affairs in which both document and query have a Boolean representation, and in which the relevance of one to the other is the parameter of interest. In addition, $\mathbf{E_{fde}}$ has a *4-valued* denotational semantics, independently developed by Belnap [4] and Dunn [18]. Compliance with the denotational approach makes this logic amenable to the various extensions (e.g. to reasoning about uncertainty) needed for modelling IR. The logic $\mathbf{E_{fde}}$ has also been investigated from the standpoint of its computational properties: while decision in the general case is co-NP-complete, whenever α and β are formulae in Conjunctive Normal Form there exists an $O(|\alpha| \cdot |\beta|)$ algorithm that tests the validity of $\alpha \rightarrow \beta$ [32].

Relevance description logics based on a 4-valued semantics have already been proposed by Patel-Schneider for use in knowledge representation, and have been proven to possess a generally better computational behaviour than their 2-valued analogues[40, 41, 42, 43]. The semantics we adopt departs from Patel-Schneider's, whose loss of inferential capabilities is too drastic for the needs of IR: in fact, that semantics sanctions the loss of *modus ponens* and, in general, of a great deal of conditional reasoning. In addition, the deduction algorithms and the completeness and complexity proofs presented by the author are rather complex, and are not modular enough to guarantee an easy adaptation to other DLs to which one might want to switch later (see Footnote 3). The 4-valued semantics for DLs that we present instead, while still adhering to the basic philosophy of relevance logics, is less restrictive, as it extends in a significant way the inferences sanctioned by the abovementioned 4-valued DLs.

3.2 The semantics

We now give the semantics of Mirlog and show, by means of examples, the differences between it and Patel-Schneider's, and between it and standard 2-valued semantics, also discussing the suitability of Mirlog for IR modelling[7]. The key difference between the 2- and the 4-valued semantics of Mirlog is that, while the former relies on the classical set of truth values $\{t, f\}$, the latter relies on its *powerset* $2^{\{t,f\}}$, which consists of the four values $\{t\}$, $\{f\}$, $\{t, f\}$ and $\{\}$. These values may be understood as representing the status of a proposition in the epistemic state of a reasoning agent. Under this view,

[7] Although we focus on a 4-valued variant of a specific DL, all our considerations on 4-valued semantics can be applied to other DLs.

if the value of a proposition contains t, then the agent has evidence to the effect – or believes – that the proposition is true. Similarly, if it contains f, then the agent has evidence to the effect that the proposition is false. The value $\{\}$ corresponds to a lack of evidence, while the value $\{t, f\}$ corresponds to the possession of contradictory evidence.

We will see that one of the effects of 4-valued semantics is the possibility of entertaining inconsistent beliefs about some proposition without this inconsistency "spreading" throughout the KB [56]. This property, that we might dub the *locality of inconsistency*, is shared by other relevance logics, and is considered one of the advantages of relevance logics, especially when modelling the epistemic states of less-than-perfect reasoning agents. The net effect in terms of IR will thus be that the presence of inconsistent beliefs about document d is unlikely to prevent a reasonable decision on whether or not to retrieve any another document.

In 4-valued semantics, an *interpretation* $\mathcal{I}$ consists of a non empty *domain* $\Delta^{\mathcal{I}}$ and of an *interpretation function* $\cdot^{\mathcal{I}}$ mapping different individuals into different elements of $\Delta^{\mathcal{I}}$, primitive concepts into functions from $\Delta^{\mathcal{I}}$ to the set $2^{\{t,f\}}$ and primitive roles into functions from $\Delta_{\mathcal{I}} \times \Delta_{\mathcal{I}}$ to $2^{\{t,f\}}$. If $\mathcal{I}$ is an interpretation, we define the *positive extension* of a concept C in $\mathcal{I}$ (written $C^{\mathcal{I}}_{+}$) as the set $\{o \in \Delta^{\mathcal{I}} : t \in C^{\mathcal{I}}(o)\}$, and the *negative extension* of a concept C in $\mathcal{I}$ (written $C^{\mathcal{I}}_{-}$) as the set $\{o \in \Delta^{\mathcal{I}} : f \in C^{\mathcal{I}}(o)\}$; the positive and negative extensions of roles are defined similarly. The positive extension of a concept C may be naturally interpreted as consisting of those domain objects that are known to be instances of the concept, while its negative extension may be likewise interpreted as consisting of those domain objects that are known *not* to be instances of the concept. Domain objects that are members of neither set are, intuitively, those neither known to be, nor known not to be instances of the concept; this is perfectly reasonable for a system that is not a perfect reasoner or does not have complete information. As for objects that are members of both sets, the intuition is that there is evidence to indicate that they are instances of the concept and, at the same time, that they are not; that is, there is inconsistent information about these objects. The semantics of a concept (or role) can then be understood as the combination of its positive extension and its negative extension. Note that, while in standard 2-valued semantics we have $C^{\mathcal{I}}_{+} \cap C^{\mathcal{I}}_{-} = \emptyset$ and $C^{\mathcal{I}}_{+} \cup C^{\mathcal{I}}_{-} = \Delta^{\mathcal{I}}$, this need not be the case with our 4-valued semantics.

As in the 2-valued case, the extensions of concepts and roles have to meet certain restrictions mirroring the informal meaning of operators. For example, the positive extension of the concept $C_1 \sqcap C_2$ must be the intersection of the positive extensions of C_1 and C_2, and its negative extension must be the *union* of their negative extensions. The complete lists of restrictions is the following:

$$
\begin{array}{rcl}
t \in \top^{\mathcal{I}}(o) & & \text{for all } o \\
f \in \top^{\mathcal{I}}(o) & & \text{for no } o \\
t \in \bot^{\mathcal{I}}(o) & & \text{for no } o \\
f \in \bot^{\mathcal{I}}(o) & & \text{for all } o \\
t \in (C_1 \sqcap C_2)^{\mathcal{I}}(o) & \text{iff} & t \in {C_1}^{\mathcal{I}}(o) \text{ and } t \in {C_2}^{\mathcal{I}}(o) \\
f \in (C_1 \sqcap C_2)^{\mathcal{I}}(o) & \text{iff} & f \in {C_1}^{\mathcal{I}}(o) \text{ or } f \in {C_2}^{\mathcal{I}}(o) \\
t \in (C_1 \sqcup C_2)^{\mathcal{I}}(o) & \text{iff} & t \in {C_1}^{\mathcal{I}}(o) \text{ or } t \in {C_2}^{\mathcal{I}}(o) \\
f \in (C_1 \sqcup C_2)^{\mathcal{I}}(o) & \text{iff} & f \in {C_1}^{\mathcal{I}}(o) \text{ and } f \in {C_2}^{\mathcal{I}}(o) \\
t \in (\neg C)^{\mathcal{I}}(o) & \text{iff} & f \in C^{\mathcal{I}}(o) \\
f \in (\neg C)^{\mathcal{I}}(o) & \text{iff} & t \in C^{\mathcal{I}}(o) \\
t \in (\forall R.C)^{\mathcal{I}}(o) & \text{iff} & \forall\ o' \in \Delta^{\mathcal{I}}, \text{if } t \in R^{\mathcal{I}}(o,o') \text{ then } t \in C^{\mathcal{I}}(o') \\
f \in (\forall R.C)^{\mathcal{I}}(o) & \text{iff} & \exists\ o' \in \Delta^{\mathcal{I}}, t \in R^{\mathcal{I}}(o,o') \text{ and } f \in C^{\mathcal{I}}(o') \\
t \in (\exists R.C)^{\mathcal{I}}(o) & \text{iff} & \exists\ o' \in \Delta^{\mathcal{I}}, t \in R^{\mathcal{I}}(o,o') \text{ and } t \in C^{\mathcal{I}}(o') \\
f \in (\exists R.C)^{\mathcal{I}}(o) & \text{iff} & \forall\ o' \in \Delta^{\mathcal{I}}, \text{if } t \in R^{\mathcal{I}}(o,o') \text{ then } f \in C^{\mathcal{I}}(o') \\
t \in (\neg P)^{\mathcal{I}}(o,o') & \text{iff} & f \in P^{\mathcal{I}}(o,o') \\
f \in (\neg P)^{\mathcal{I}}(o,o') & \text{iff} & t \in P^{\mathcal{I}}(o,o')
\end{array}
$$

In the 4-valued case, the notion of an interpretation $\mathcal{I}$ satisfying an assertion or a definition relies only on the *positive* extensions of the concepts and roles involved[8], and is thus basically unchanged with respect to the one we have given for the 2-valued case. Formally, the definition $C_1 \sqsubseteq C_2$ (resp. $R_1 \sqsubseteq R_2$) is *satisfied* by an interpretation $\mathcal{I}$ iff ${C_1}^{\mathcal{I}}_{+}(o)$ implies ${C_2}^{\mathcal{I}}_{+}(o)$ for all $o \in \Delta_{\mathcal{I}}$ (resp. ${R_1}^{\mathcal{I}}_{+}(o,o')$ implies ${R_2}^{\mathcal{I}}_{+}(o,o')$ for all $o, o' \in \Delta_{\mathcal{I}}$). An interpretation $\mathcal{I}$ *satisfies an assertion* α iff $t \in C^{\mathcal{I}}(\gamma(a))$ in case $\alpha = C[a]$ or $t \in R^{\mathcal{I}}(\gamma(a_1), \gamma(a_2))$ in case $\alpha = R[a_1, a_2]$. We will also say that $\mathcal{I}$ *f-satisfies an assertion* α iff $f \in C^{\mathcal{I}}(\gamma(a))$ in case $\alpha = C[a]$, whereas $f \in R^{\mathcal{I}}(\gamma(a_1), \gamma(a_2))$ in case $\alpha = R[a_1, a_2]$. Satisfiability is extended to assertional formulae as follows.

Definition 1 *Let $\mathcal{I}$ be an interpretation.*

1. $\mathcal{I}$ satisfies *an assertional formula* $\gamma_1 \wedge \gamma_2$ *iff it satisfies both* γ_1 *and* γ_2*;*
2. $\mathcal{I}$ f-satisfies *an assertional formula* $\gamma_1 \wedge \gamma_2$ *iff it f-satisfies* γ_1 *or* $\mathcal{I}$ *f-satisfies* γ_2*;*
3. $\mathcal{I}$ satisfies *an assertional formula* $\gamma_1 \vee \gamma_2$ *iff it satisfies* γ_1 *or* $\mathcal{I}$ *satisfies* γ_2*;*
4. $\mathcal{I}$ f-satisfies *an assertional formula* $\gamma_1 \vee \gamma_2$ *iff it f-satisfies both* γ_1 *and* γ_2*;*
5. $\mathcal{I}$ satisfies *an assertional formula* $\sim \gamma$ *iff it f-satisfies* γ*;*
6. $\mathcal{I}$ f-satisfies *an assertional formula* $\sim \gamma$ *iff it satisfies* γ*.*

Given two Mirlog concepts C_1 and C_2, C_1 *is subsumed by* C_2 (written $C_1 \sqsubseteq_4 C_2$) iff ${C_1}^{\mathcal{I}}_{+} \subseteq {C_2}^{\mathcal{I}}_{+}$ for every interpretation $\mathcal{I}$, and C_1 is *equivalent* to C_2 (written $C_1 \equiv_4 C_2$) iff ${C_1}^{\mathcal{I}}_{+} = {C_2}^{\mathcal{I}}_{+}$ for every interpretation $\mathcal{I}$. $\sqsubseteq_4$ and $\equiv_4$ are extended to roles in a straightforward way. Finally, a KB Σ entails an assertion α ($\Sigma \models_4 \alpha$) iff every interpretation satisfying the former also satisfies the latter.

[8] For the motivations underlying this choice, see the discussion on *t-entailment* in [40].

3.3 Soundness and incompleteness

One important property of Mirlog is that reasoning in it is *sound* with respect to classical semantics; that is, every inference that can be drawn within Mirlog can also be drawn within its corresponding 2-valued logic presented, for illustrative purposes, in the previous section. This means that a user acquainted with classical semantics does not run the risk of being offered a conclusion she would not subscribe to.

In order to show this, it suffices to notice that the set of 2-valued interpretations is a (proper) subset of the set of 4-valued interpretations. Consider in fact a 4-valued interpretation $\mathcal{I}$ such that the positive and negative extensions of every primitive concept A and primitive role P are both disjoint and exhaustive, i.e. $A^{\mathcal{I}}_{-} = \Delta^{\mathcal{I}} \setminus A^{\mathcal{I}}_{+}$ and $P^{\mathcal{I}}_{-} = (\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}) \setminus P^{\mathcal{I}}_{+}$. By a case analysis on the semantics of the operators, it can be seen that such an interpretation is a 2-valued interpretation for DLs; in fact, note that in these interpretations, given any concept C and any role R, $t \in C^{\mathcal{I}}(o)$ iff $f \notin C^{\mathcal{I}}(o)$ and $t \in R^{\mathcal{I}}(o, o')$ iff $f \notin R^{\mathcal{I}}(o, o')$,

The soundness result follows immediately, since inferring e.g. an assertion $C[a]$ from a KB Σ corresponds to checking if Σ entails $C[a]$, hence to checking whether all interpretations satisfying Σ also satisfy $C[a]$. We then have the following:

Lemma 1 *Let Σ be a KB, α an assertion and C_1 and C_2 two concepts. Then $C_1 \preceq_4 C_2$ implies that $C_1 \preceq C_2$, and $\Sigma \models_4 \alpha$ implies that $\Sigma \models \alpha$.*

However, reasoning in Mirlog is not *complete*, i.e. not every inference that can be drawn based on 2-valued semantics can also be drawn within Mirlog. This was to be expected, as both soundness and completeness would mean that Mirlog and its correspondent 2-valued version are the same logic. Coupled with the previous Lemma, this means that the conclusions licensed by Mirlog are then a *proper* subset of those licensed by 2-valued semantics. What we want to show is that this is also an *interesting* subset for IR purposes, i.e. that the conclusions to which Mirlog renounces are somehow debatable, and that Mirlog captures, as a result, some natural intuitions about reasoning that also translate into more intuitive behaviour once applied to modelling IR. We will describe this behaviour by a number of examples.

3.4 Modus ponens on roles

Let us consider the KB Σ, shown in Figure 1, consisting of assertions about documents `doc1` and `doc2`. `doc1` is asserted to be a document with only Italian authors 2, one of which is `mario` 3. `doc1` has two components: `c11`, which is a text, and `c12`, whose medium is unknown. There is a reference from `c11` to `c12` and from the latter to `c21`, which is a video and a component of `doc2`. This latter is a multimedia document, all of which components have a string as title; `doc2` is authored by `mario`, which is asserted to be not Italian. We will use this KB as a running example throughout the remainder of this section to exemplify some points about the relationship between Mirlog's and other semantics. First of all, it is interesting to observe that the following entailment relationship holds:

$$\Sigma \models_4 (\texttt{MultimediaDoc} \sqcap \exists\texttt{Author.Italian})[\texttt{doc2}]. \tag{7}$$

That `doc2` has an Italian author follows from the fact that Mario, who is an author of `doc2` (by 4), is also an author of `doc1` (by 3), all authors of which are Italian (by 2). This example shows that assertions concerning a certain document (such as (2), which is about `doc1`) may have an influence on the retrieval of a different document (`doc2`, in our case). This form of inference is indeed desirable for IR purposes, but is not sanctioned by Patel-Schneider's 4-valued semantics, which, disallowing *modus ponens,* rules out a great deal of conditional reasoning. More generally, our semantics can be proved

$$
\begin{array}{cr}
(\texttt{Document} \sqcap \forall \texttt{Author.Italian})[\texttt{doc1}], & (2)\\
\texttt{Author}[\texttt{doc1}, \texttt{mario}], & (3)\\
\texttt{Component}[\texttt{doc1}, \texttt{c11}],\ \exists\texttt{Medium.Text}[\texttt{c11}],\ \texttt{Component}[\texttt{doc1}, \texttt{c12}], & \\
\texttt{References}[\texttt{c11}, \texttt{c12}],\ \texttt{References}[\texttt{c12}, \texttt{c21}], & \\
(\texttt{MultimediaDoc} \sqcap (\forall\texttt{Component}.\exists\texttt{Title.String}))[\texttt{doc2}], & \\
\texttt{Author}[\texttt{doc2}, \texttt{mario}], & (4)\\
\neg\texttt{Italian}[\texttt{mario}], & (5)\\
\texttt{Component}[\texttt{doc2}, \texttt{c21}],\ \exists\texttt{Medium.Video}[\texttt{c21}], & \\
\texttt{Video} \sqsubseteq \neg\texttt{Text} & (6)
\end{array}
$$

Figure 1: A Mirlog knowledge base

to license inferences conforming to the following schemata: for all concepts C_1 and C_2, roles R and individuals a_1, a_2, we have that:

$$\{(\forall R.C_1)[a_1], R[a_1, a_2]\} \models_4 C_1[a_2]$$
$$\{(\forall R.C_1)[a_1], (\exists R.C_2)[a_1]\} \models_4 (\exists R.(C_1 \sqcap C_2))[a_1]$$

We call these schemata *modus ponens on roles.* The key difference between our account and Patel-Schneider's lies in the semantics of the $\forall$ operator. Patel-Schneider's t-condition for $\forall$ is:

$$t \in (\forall R.C)^{\mathcal{I}}(o) \quad \text{iff} \quad \forall\ o' \in \Delta^{\mathcal{I}}, f \in R^{\mathcal{I}}(o, o') \text{ or } t \in C^{\mathcal{I}}(o')$$

while ours, we recall, is:

$$t \in (\forall R.C)^{\mathcal{I}}(o) \quad \text{iff} \quad \forall\ o' \in \Delta^{\mathcal{I}}, \text{if } t \in R^{\mathcal{I}}(o, o') \text{ then } t \in C^{\mathcal{I}}(o')$$

Now, it can be verified that, according to Patel-Schneider rules, there exists a model $\mathcal{I}$ of Σ such that both t and f are in $\texttt{Author}^{\mathcal{I}}(\texttt{doc1}^{\mathcal{I}}, \texttt{mario}^{\mathcal{I}})$, and such that $t \notin \texttt{Italian}^{\mathcal{I}}(\texttt{mario}^{\mathcal{I}})$; it immediately follows that Σ does not entail $\texttt{Italian}[\texttt{mario}]$ and *a fortiori* it does not entail $(\texttt{MultimediaDoc} \sqcap \exists\texttt{Author.Italian})[\texttt{doc2}]$.

3.5 Inconsistent knowledge bases and tautologous queries

The preceding discussion has focussed on showing what inferences *can* be drawn in our 4-valued semantics (and, in some cases, cannot according to other brands of 4-valued semantics). Now we concentrate on inferences that are valid in the standard 2-valued semantics but are not licensed by our semantics. The examples of this section are reminiscent of the so-called "paradoxes of logical implication" for classical logic (or rather, of a DL version of them).

First, note that the KB Σ, according to classical terminology, is inconsistent: indeed, as already argued, assertions (2) and (3) provide implicit evidence to the fact that Mario is Italian, while assertion (5) explicitly states that he is not. As a consequence, both the following hold:

$$\Sigma \models_4 \neg\texttt{Italian}[\texttt{Mario}] \tag{8}$$
$$\Sigma \models_4 \texttt{Italian}[\texttt{Mario}]. \tag{9}$$

The same holds in 2-valued semantics, because of the soundness of entailment. Moreover, in 2-valued semantics any assertion follows from Σ, due to its inconsistency. So, for example:

$$\Sigma \models ((\exists \texttt{Medium}.\texttt{Video}) \sqcap (\exists \texttt{Author}.\texttt{Italian}))[\texttt{c21}],$$

which means that c21 would be retrieved in the response to a query asking for videos with an Italian author. However, this retrieval falls short of relevance, as there is nothing in Σ supporting the Italianhood of c21's authors. Therefore, a model aiming at capturing relevance should forbid retrievals like this, and this is in fact what entailment does; that is, as it can be easily verified:

$$\Sigma \not\models_4 ((\exists \texttt{Medium}.\texttt{Video}) \sqcap (\exists \texttt{Author}.\texttt{Italian}))[\texttt{c21}]. \qquad (10)$$

This example shows a fundamental advantage of a semantics inspired to relevance: KBs that are inconsistent from a 2-valued semantics point of view, do not entail every assertion, or, put in another way, the effect of inconsistencies is localized, as the following inferences show:

$$\begin{array}{lll} \Sigma & \models_4 & (\exists \texttt{Author}.\texttt{Italian})[\texttt{doc2}] \\ \Sigma & \models_4 & (\exists \texttt{Author}.\neg\texttt{Italian})[\texttt{doc2}] \end{array}$$

Dually, assertions based on concepts whose extension is, in 2-valued semantics, always the entire domain of an interpretation, and which therefore closely resemble *tautologies,* are not necessarily entailed by every KB. For instance, in 2-valued semantics any document component is either of type video or of a type different from video, a fact formally captured by the following (true) implication relation:

$$\Sigma \models (\forall \texttt{Medium}.(\texttt{Video} \sqcup \neg\texttt{Video}))[\texttt{c12}].$$

However, Σ says nothing about the medium of c12, thus, strictly speaking, there is no relevance relation between c12 as described in Σ and the query $(\forall \texttt{Medium}.(\texttt{Video} \sqcup \neg\texttt{Video}))$. And in fact:

$$\Sigma \not\models_4 (\forall \texttt{Medium}.(\texttt{Video} \sqcup \neg\texttt{Video}))[\texttt{c12}] \qquad (11)$$

holds. To see why, note that there is a model $\mathcal{I}$ of Σ such that for some $o' \in \Delta^{\mathcal{I}}$, $t \in \texttt{Medium}^{\mathcal{I}}(\texttt{c12}^{\mathcal{I}}, o')$, and $\texttt{Video}^{\mathcal{I}}(o') = \{\}$.

Cases of "inconsistent" KBs or cases of "tautologous" queries[9] have been deemed of debatable importance to IR. However, while in general the intuitive behaviour of our logic also in these "extreme cases" is a witness of its quality, we argue that inconsistencies in document bases are going to be the rule more than the exception in light of the globalization process that the retrieval of information is more and more experiencing. As far as tautologous queries is concerned, the mechanism that prevents their inference has an importance that goes much beyond such queries, as illustrated by next section.

3.6 Reasoning by cases

The behaviour of relevance semantics on tautologies has an impact on the inference of assertions that are not tautologous by themselves, but which somehow require the establishment of a tautology for their deduction. This pattern occurs in a reasoning scheme termed *reasoning by cases*, which the following exemplifies. Let us consider the query α given by:

$$(\exists \texttt{Component}.(\exists \texttt{Medium}.\texttt{Text} \sqcap \exists \texttt{References}.\exists \texttt{Medium}.\neg\texttt{Text}))$$

[9] Quotes are meant to remark that these words should be understood in their 2-valued reading.

on the KB Σ already introduced. We want to check whether doc1 should be retrieved in response to this query, *i.e.* whether $\Sigma \models_4 \alpha[\texttt{doc1}]$. Let $\mathcal{I}$ be the 4-valued model of Σ introduced above, taking no position on the medium of c12. By straightforward semantical arguments, it may be seen that

$$t \notin (\exists\texttt{Component}.(\exists\texttt{Medium}.\texttt{Text} \sqcap \exists\texttt{References}.\exists\texttt{Medium}.\neg\texttt{Text}))^{\mathcal{I}}(\texttt{doc1}^{\mathcal{I}}),$$

from which it follows that $\Sigma \not\models_4 \alpha[\texttt{doc1}]$. However, perhaps surprisingly, $\Sigma \models \alpha[\texttt{doc1}]$. At first, it would seem that this is not the case, since c11 and c12 are the only known components of doc1, and neither of them seems to be a text which references a document of a different medium. But, let us consider a 2-valued model $\mathcal{J}$ of Σ and let us *reason by cases.* $\mathcal{J}$, unlike $\mathcal{I}$ above, must support either the truth of $(\exists\texttt{Medium}.\neg\texttt{Text})[\texttt{c12}]$ or its falsity. And this is enough for the inference to hold. For in the former case, doc1 has as a component c11, which is a text and referencing c12, a non-text. In the latter case, doc1 has as a component c12, which is a text referencing c21, a video and therefore, by definition 6, a non-text. In both cases α is true in $\mathcal{J}$, and by generalization $\Sigma \models \alpha[\texttt{doc1}]$.

3.7 Conclusions

To sum up, what kind of relevance relation is captured by $\models_4$?

A first answer is that, roughly speaking, a KB Σ entails everything that is in the transitive closure of Σ by means of *modus ponens* on roles and the operators $\sqcap, \sqcup, \neg, \exists$, as (7), (8) and (9) demonstrate. All other inferences are left out, as (10), (11) and the example on reasoning by cases show. More precisely, in order for $\Sigma \models_4 \alpha$ to hold, the structural components of α must have an analogue in Σ, modulo *modus ponens* on roles.

In less technical terms, a KB Σ entails everything that is *explicitly supported* or, we might say, everything for which there are *relevant premises.* The inference modelled by our semantics can thus be seen akin to what has been termed *shallow reasoning* in the AI literature, *i.e.* a mode of reasoning in which a logical agent only draws quick inferences that do not overtax its resources. Those inferences that 2-valued semantics licenses and 4-valued semantics does not are those for drawing which the agent must reason, as Levesque says [33], in *puzzle mode*, i.e. in the style that humans adopt once we try to solve a challenging mathematical problem or a logical puzzle. This interpretation brings further evidence to the fact that Mirlog is an adequate tool for IR, whose reasoning task seems quite different from that of a working mathematician.

4 Closures

In the preceding sections we have discussed in detail the issue of how to deal appropriately with document (and query) *content.* We now turn our attention to the representation of those document features that require, upon retrieval, *closed-world* reasoning. One important class of such features concerns document *structure,* which is an important issue for retrieval, as queries can make explicit references to the composition of documents to be retrieved, e.g. by requesting documents that deal with a particular topic *and* contain photographs relating to this topic plus coordinated text. Another important class is document *profile,* which includes knowledge about the external characterization of a document, such as its title, authors, producing date, copyrights and the like.

It so happens that the language of DLs is essentially adequate for the representation of the features in question, while their inferential apparatus is not. This is due to the fact that, for instance, when reasoning about structure, it is both convenient and adequate to equate what is *known* about a document with what is *true* about it. This point, which we now argue in full detail, leads to the requirement that certain reasoning be informed by the closed-world assumption, and in order to specify exactly when to adopt this assumption, we introduce a new operator that allows to distinguish items of knowledge liable of a closed-world reading from the rest.

4.1 An informal introduction to closure assertions

Let us consider the KB Σ presented in Figure 2, containing structural and profile information about two documents, doc3 and doc4. About the former, Σ knows that it is a letter sent by a Scottish man named Guglielmo. About the latter, all Σ knows is that it is a book. We further suppose that the knowledge in Σ is also all there is to know about the two documents, all the rest being not true of them; thus, for instance, d has no sender other than Guglielmo.

$$\texttt{Letter[doc3]}\ ,\ \texttt{Sender[doc3,Guglielmo]}\ ,\ \texttt{Scottish[Guglielmo]}\ ,\ \texttt{Book[doc4]}$$

Figure 2: A Mirlog knowledge base.

Because of the nature of the knowledge held by Σ, one would like to have doc3 retrieved in response to the query $\neg$Book, asking for all individuals that are not books. However,

$$\Sigma \quad \not\models \quad \neg\texttt{Book[doc3]},$$

as there are 2-valued models of Σ in which the individual named doc3 is both in the extension Letter and Book, so making $\neg$Book[doc3] false. The corresponding 4-valued models support: $\Sigma \not\models_4$ $\neg$Book[doc3], hence doc3 would *not* be retrieved in our model as presented so far.

In order to solve this problem, one could add the definition: Letter$\sqsubseteq$ $\neg$Book to Σ; however, this definition would introduce an inconsistency for all letters that are also published as books. The relevance semantics of Mirlog would prevent these inconsistencies from breaking the whole KB, but they would anyway operate at the local level, a somewhat disturbing fact. In addition, there are other inferences that, on the basis of the same intuition, one would like to draw from Σ and that are not dealt with by definitions. One of these inferences is that all doc3's senders are Scottish. But, again, $\Sigma \not\models \forall$Sender.Scottish[doc3], and, *a fortiori,* $\Sigma \not\models_4 \forall$Sender.Scottish[doc3].

A radical solution to the problem would be to embed in Σ a complete description of doc3. Such description would consist of all positive assertions about doc3, plus:

- one concept assertion of the form $\neg A$[doc3] for all primitive concepts A which doc3 is not an instance of, and
- one role assertion of the form $\neg R$[doc3,c] for all primitive roles R and individuals c such that R[d,c]] is not the case.

Given that the catalog of a realistic document base is likely to comprise at least hundreds of concepts and roles, and thousands of individuals, the complete description of doc3 would require an overwhelming amount of assertions.

Our solution to this problem is to extend the IR model developed so far with (meta-)assertions on elements of the language that force a closed-world interpretation of the (normal) assertions concerning such elements. For instance, a closure assertion on the individual a, would mean that the KB contains, whether explicitly or implicitly, everything that is true about a, and every other fact concerning a is to be considered as false. A meta-assertion like the above is called a *closure assertion*, as it induces a reading of the information concerning a clearly reminiscent of the *closed-world assumption* used e.g. in logic programming and deductive databases. The individuals that are the subject of closure assertions are said to be *closed.*

The information provided by closure assertions must guide the inferential behaviour of the system on closed individuals in a way that reflects intuition. More precisely, while the lack of information on the truth of a fact concerning a non-closed individual is to be interpreted in the usual way, *i.e.* as lack of knowledge about the given fact *and* about its negation, when a closed individuals is involved this is to be interpreted as knowledge of the negation of the given fact. Returning to the previous example, the intended interpretation of closure assertions would grant the following inferences:

$$\begin{array}{lcl} \Sigma \cup \{\mathtt{CL(doc3)}\} & \models_c & \neg\mathtt{Book}[\mathtt{doc3}] \\ \Sigma \cup \{\mathtt{CL(doc3)}\} & \models_c & \forall\mathtt{Sender.Scottish}[\mathtt{doc3}] \end{array}$$

where $\models_c$ is the inference relation of the new logic. The relation $\models_c$ should clearly be non-monotonic, i.e. the addition of new information might possibly block inferences that were previously valid. For instance, the following should hold:

$$\Sigma \cup \{\mathtt{CL(doc3)}\} \cup \{\mathtt{Book}[\mathtt{doc3}]\} \quad \not\models_c \quad \neg\mathtt{Book}[\mathtt{doc3}].$$

4.2 Relation to other approaches

Since the seminal paper by Reiter [44], many forms of closed-world assumption (CWA) have been investigated (see [34, Chapter 7] for a thorough review). The proposal most similar in spirit to ours is the so-called *careful CWA* [21], by means of which one can confine the closed-world reading to a pre-specified subset of predicate symbols only. Without going into the details of this and the other CWA proposals, we observe that neither careful CWA nor other forms of CWA seem suited to our program of allowing the closed-world reading to be applied selectively to either pre-specified predicate symbols or individuals. In fact:

- careful CWA does not allow the restriction of the CWA to specified *individuals*;
- every form of CWA can operate only on KBs that are *universal theories without equality.* Notoriously, a MIRLOG KB is not in general a universal theory[10].

Versions of the CWA specifically formulated for DLs have recently appeared [14, 15] which are based on the use (within the query language) of an explicit epistemic operator **K**, whose natural language reading is the adjective "known". The basic idea behind these proposals is to enforce a CWA reading of the information about an individual a by using the operator **K** when checking whether a given fact about a is entailed by the KB. Applied to the previous example, this means that in order to obtain a positive answer on the membership of doc3 to the ¬Book concept, one has to check whether ¬KBook[doc3] ("d is not a known book") is entailed by Σ, *i.e.* ask whether d *is not known* by the KB to be a book, which indeed turns out to be the case. Analogously, checking whether ∀**K**Sender.Scottish[doc3] ("all known senders of doc3 are Scottish") is entailed by Σ returns a positive answer, because there is only one *known* sender of d and he happens to be Scottish.

As made clear by these examples, the use of an epistemic operator in queries would allow one to ask questions not only about the world, but also about the state of knowledge of the KB [45]. It is by now evident that this use permits to capture, among other things, some form of CWA. However, clear connections between epistemic queries posed to DL KBs and the various CWA formulations have not been established yet, except for a very restricted case (see Theorem 5.1 in [14]). Thus, strictly speaking, one cannot claim full control of how epistemic queries to DL KBs realize CWA.

[10] A universal theory of first order logic is a set of formulae such that their prenex normal form does not contain existential quantifiers. The simple MIRLOG theory $\{\exists R_1.(\forall R_2.C)[a]\}$ is equivalent to the FOL theory $\{\exists x \exists y.(R_1(a,x) \wedge R_2(x,y) \wedge C(y))\}$, which is in prenex normal form but is not universal.

Besides this formal argument, there is a methodological reason why the adoption of the epistemic approach in our IR setting is problematic. Let us consider the KB $\Sigma_1 = \{\texttt{Letter}[\texttt{d}], \texttt{CL}(\texttt{d}), \texttt{Letter}[\texttt{a}]\}$ and the query $\alpha = \neg\texttt{Book}$. According to our intended meaning of closure assertions, the answer to α in Σ_1 should be the set $\{\texttt{d}\}$. In order to obtain the same behaviour by means of epistemic queries, α should be broken down (behind the scene) into two queries $\alpha_1 = \neg\texttt{Book}[\texttt{a}]$ and $\alpha_2 = \neg\mathbf{K}\texttt{Book}[\texttt{d}]$. In order to perform this transformation, the underlying IR system must be told which individuals are closed. But then, once the explicit specification of closed individuals is available, it is preferable to use it in the most direct and neat way, *i.e.* by devising a semantics that reflects the intuition behind these assertions. And this is precisely our approach. Furthermore, it is not at all clear how the closure of a role for a certain individual, a feature of CLASSIC [6] that is offered by our model under the name of pointwise role closure, would be simulated in the epistemic approach.

4.3 Knowledge bases with closures

Let a be an individual, P a primitive role and T a primitive term. Then:

- An *individual closure* is an expression of type $\texttt{CL}(a)$. The individual a is said to be *closed.*
- A *primitive closure* is an expression of type $\texttt{CL}(T)$. The term T is said to be *closed.*
- A *pointwise role closure* is an expression of type $\texttt{CL}(a, P)$. The individual a is said to be *closed w.r.t. role P*.

A *CBox* is a finite set of closures. An MIRLOG KB is extended to be a pair $\langle\Sigma, \Omega\rangle$, where Σ is a set of assertional formulae and definitions, and Ω is a CBox. Note that, since $A \sqsubseteq C$ and $C \sqsubseteq A$ define concept A to be equivalent to C, $\texttt{CL}(A)$closes the concept C. Hence, closures of complex concepts (and roles) are allowed in MIRLOG.

After presenting syntax, we now discuss the semantics of closures.

The first, important semantic shift required by closures is the introduction of a fixed domain of interpretation, necessary to properly deal with trans-world identity of individuals. This shift is obtained by replacing the notion of interpretation by that of *c-interpretation,* defined in the following.

Let Δ be the *domain,* a countable infinite set of symbols, called *parameters* (denoted by p and p') and γ a fixed injective function from $\mathcal{O}$ to Δ. A *c-interpretation* $\mathcal{I}$ is a 4-valued interpretation such that:

1. $\Delta^{\mathcal{I}} = \Delta$ and
2. for all individuals a, $a^{\mathcal{I}} = \gamma(a)$.

The notion of satisfaction of normal assertions is extended to c-interpretations in the obvious way. Unless otherwise specified, in the following by "interpretation" we mean "c-interpretation". With $\mathcal{M}(\Sigma)$ we indicate the set of all (4-valued) models of Σ.

Satisfaction of closures is defined on the basis of a notion of minimal knowledge, modelled by epistemic interpretations. An *epistemic interpretation* is a pair $\langle\mathcal{I}, \mathcal{W}\rangle$ where $\mathcal{I}$ is an interpretation and $\mathcal{W}$ is a set of interpretations.

Definition 2 *An epistemic interpretation $\langle\mathcal{I}, \mathcal{W}\rangle$* satisfies *a closure* $\texttt{CL}(a)$*if and only if the following conditions hold:*

1. *for every primitive concept symbol A, $t \in A^{\mathcal{I}}(\gamma(a))$ iff $t \in A^{\mathcal{J}}(\gamma(a))$ for all $\mathcal{J} \in \mathcal{W}$;*
2. *for every primitive concept symbol A, $f \in A^{\mathcal{I}}(\gamma(a))$ iff $t \notin A^{\mathcal{J}}(\gamma(a))$ for some $\mathcal{J} \in \mathcal{W}$;*

3. *for every primitive role symbol P and parameter $p \in \Delta$, $t \in P^{\mathcal{I}}(\gamma(a), p)$ iff $t \in P^{\mathcal{J}}(\gamma(a), p)$ for all $\mathcal{J} \in \mathcal{W}$;*

4. *for every primitive role symbol P and parameter $p \in \Delta$, $f \in P^{\mathcal{I}}(\gamma(a), p)$ iff $t \notin P^{\mathcal{J}}(\gamma(a), p)$ for some $\mathcal{J} \in \mathcal{W}$.* ∎

In words, for any model of a KB $\langle \Sigma, \Omega \rangle$ and closed individual a, $a^{\mathcal{I}}$ is allowed in the positive extension of a primitive concept A just in case $A(a)$ is entailed by Σ, in symbols $\Sigma \models_4 A(a)$. As a consequence, the lack of positive information will allow us, as will be soon shown, to infer the corresponding negative information. Similarly for roles. The semantics of primitive closures is perfectly dual; it constrains the extensions of closed primitive concepts and roles with respect to parameters.

Definition 3 *Let A be a primitive concept. An epistemic interpretation $\langle \mathcal{I}, \mathcal{W} \rangle$* satisfies *a closure* $\mathtt{CL}(A)$*if and only if the following conditions hold:*

1. *for every $p \in \Delta$, $t \in A^{\mathcal{I}}(p)$ iff $t \in A^{\mathcal{J}}(p)$ for all $\mathcal{J} \in \mathcal{W}$;*

2. *for every $p \in \Delta$, $f \in A^{\mathcal{I}}(p)$ iff $t \notin A^{\mathcal{J}}(p)$ for some $\mathcal{J} \in \mathcal{W}$.*

An epistemic interpretation satisfies *a closure* $\mathtt{CL}(P)$*, where P is a primitive role, if and only if the following conditions hold:*

3. *for all $p, p' \in \Delta$, $t \in P^{\mathcal{I}}(p, p')$ iff $t \in P^{\mathcal{J}}(p, p')$ for all $\mathcal{J} \in \mathcal{W}$;*

4. *for all $p, p' \in \Delta$, $f \in P^{\mathcal{I}}(p, p')$ iff $t \notin P^{\mathcal{J}}(p, p')$ for some $\mathcal{J} \in \mathcal{W}$.* ∎

Finally, we observe that the pointwise closure $\mathtt{CL}(a, P)$is equivalent to the assertions $(\forall P.A_p)(a)$ and $\mathtt{CL}(A_p)$, where A_p is a new primitive concept. We will therefore understand the semantics of pointwise closures in terms of that of primitive closures, and concentrate, from now on, only on individual and primitive closures.

An epistemic interpretation *satisfies* (is a *model* of) a set of closures if and only if it satisfies each closure in the set.

Definition 4 *Let $\langle \Sigma, \Omega \rangle$ be a KB. An interpretation $\mathcal{I}$* satisfies *(is a* model *of) $\langle \Sigma, \Omega \rangle$ if and only if $\mathcal{I}$ is a model of Σ and $\langle \mathcal{I}, \mathcal{M}(\Sigma) \rangle$ is a model of Ω.* ∎

Essentially, in order to be a model of a KB, an interpretation has to satisfy the "normal" assertions in Σ and the requirements imposed by closures, given in the previous definitions. Finally,

Definition 5 *A KB $\langle \Sigma, \Omega \rangle$* c-entails *a query Q, written $\langle \Sigma, \Omega \rangle \models_4^c Q$, if and only if all models of $\langle \Sigma, \Omega \rangle$ satisfy Q.* ∎

4.4 Properties of closures

Let us consider the KB $\langle \Sigma, \Omega \rangle$ where Σ is the set of assertions shown in Figure 2, and:

$$\Omega = \{\mathtt{CL}(\mathtt{doc3})\}.$$

Thanks to the closure of `doc3`, in all the models of $\langle \Sigma, \Omega \rangle$, $\mathtt{doc3}^{\mathcal{I}}$ only belongs to the positive extension of `Letter`, that is $t \in \mathtt{Letter}^{\mathcal{I}}(\mathtt{doc3}^{\mathcal{I}})$ and $t \notin A^{\mathcal{I}}(\mathtt{doc3}^{\mathcal{I}})$ for all other primitive concepts A. By rule 2 of definition 2, this means that in all the models of $\langle \Sigma, \Omega \rangle$, $f \in \mathtt{Book}^{\mathcal{I}}(\mathtt{doc3}^{\mathcal{I}})$, therefore, as desired:

$$\langle \Sigma, \Omega \rangle \models_4^c \neg\mathtt{Book}[\mathtt{doc3}].$$

For the same reason, in all the models of $\langle\Sigma,\Omega\rangle$, the positive extension of $\mathtt{Sender}$ is given by:

$$\mathtt{Sender}_+^{\mathcal{I}} \;=\; \{\langle \mathtt{doc3}^{\mathcal{I}}, \mathtt{Guglielmo}^{\mathcal{I}}\rangle\}$$

Because in all such models $\mathtt{Guglielmo}^{\mathcal{I}}$ is in the extension of $\mathtt{Scottish}$, again as desired:

$$\langle\Sigma,\Omega\rangle \models_4^c \forall\mathtt{Sender}.\mathtt{Scottish}[\mathsf{doc3}].$$

This latter inference could also be obtained by closing the role $\mathtt{Sender}$, *i.e.* by having $\mathtt{CL}(\mathtt{Sender})$ in Ω.

A formal investigation of the features of closures follows.

We begin by illustrating a close relationship between completely closed KBs w.r.t. individuals and completely closed KBs w.r.t. primitives.

Proposition 1 *Let Σ be an ABox, let $C(a)$ an assertion, let Ω_1 be such that all individuals in Σ are closed and such that a is closed, let Ω_2 be such that all primitives in Σ are closed and such that all primitives in C are closed, then $\langle\Sigma,\Omega_1\rangle \models_4^c C(a)$ iff $\langle\Sigma,\Omega_2\rangle \models_4^c C(a)$.* ∎

As a consequence, all theorems for completely closed KBs w.r.t. individuals are easily adaptable to KBs completely closed w.r.t. primitives.

A concept C is said to be *quantifier free* if no quantifier occurs in it. Moreover, a KB is called:

- *completely closed w.r.t. individuals* iff all individuals appearing in it are closed;
- *completely closed w.r.t. primitives* iff all primitives appearing in it are closed;
- *completely closed* iff both previous conditions hold.

In classical logic, a theory is said to be complete if, for any sentence α, either α or its negation follows from the theory. The next two theorems show that closing an individual or a primitive amounts to make the knowledge about it complete in the classical sense. Since an assertion containing a quantifier involves also other individuals, a *proviso* is required in the first part of the next theorem. The second part shows that, when all the individuals are closed, the *proviso* is no longer needed.

Proposition 2 *Let $\langle\Sigma,\Omega\rangle$ be a KB, $\mathtt{CL}(a) \in \Omega$, and $C(a)$ a concept assertion. Then:*

1. *either $\langle\Sigma,\Omega\rangle \models_4^c C(a)$ or $\langle\Sigma,\Omega\rangle \models_4^c \neg C(a)$, for any quantifier free C;*
2. *if $\langle\Sigma,\Omega\rangle$ is completely closed w.r.t. individuals, then either $\langle\Sigma,\Omega\rangle \models_4^c C(a)$ or $\langle\Sigma,\Omega\rangle \models_4^c \neg C(a)$, for any C.* ∎

For closed terms we have:

Proposition 3 *Let $\langle\Sigma,\Omega\rangle$ be a KB. Then if $\mathtt{CL}(A) \in \Omega$ then for all individuals a either $\langle\Sigma,\Omega\rangle \models_4^c A(a)$ or $\langle\Sigma,\Omega\rangle \models_4^c \neg A(a)$.* ∎

It is natural to ask how c-entailment relates to entailment. The answer to this question comes in three steps. First, a KB with no closures is equivalent to (*i.e.* has the same models as) a set of normal assertions; this means that c-entailment coincides with entailment on closure-less KBs.

Proposition 4 *Let Σ be a set of assertions. Then an interpretation is a model of $(\Sigma,\emptyset)$ iff it is a model of Σ.* ∎

Second, when closures are considered, c-entailment extends entailment, that is $\models_4 \subset \models_4^c$.

Proposition 5 *Let $\langle\Sigma, \Omega\rangle$ be a KB and $C(a)$ an assertion. Then $\Sigma \models_4 C(a)$ implies $\langle\Sigma, \Omega\rangle \models_4^c C(a)$.* ∎

In order to show that $\models_4 \neq \models_4^c$, it suffices to consider the example completed at the beginning of this Section. As we have seen, $\Sigma \not\models_4 \neg\texttt{Book}[\texttt{doc3}]$, whereas $\langle\Sigma, \Omega\rangle \models_4^c \neg\texttt{Book}[\texttt{doc3}]$.

Third, c-entailment captures a form of Closed-World Assumption (CWA): a positive assertion is c-entailed if it is entailed, while a negative assertion is c-entailed if the corresponding positive assertion is not entailed. Also the converse holds, provided that the KB is satisfiable, because, as it follows from the semantics of closures, a closed individual can only be associated with the classical truth values ($\{t\}$ and $\{f\}$), hence on closed terms the KB behaves as a classical theory (as we will see in the next section, this has an impact on inconsistency). The next theorem formalizes this fact, showing exactly what is the inferential gain of c-entailment over classical entailment.

Proposition 6 *Let $\langle\Sigma, \Omega\rangle$ be a KB. Then*

1. *if $\texttt{CL}(a) \in \Omega$ then for each primitive concept A,*

   (a) $\Sigma \models_4 A(a)$ implies $\langle\Sigma, \Omega\rangle \models_4^c A(a)$;

   (b) $\Sigma \not\models_4 A(a)$ implies $\langle\Sigma, \Omega\rangle \models_4^c \neg A(a)$.

   Conversely, if $\langle\Sigma, \Omega\rangle$ is satisfiable, then for each primitive concept A,

   (c) $\langle\Sigma, \Omega\rangle \models_4^c A(a)$ implies $\Sigma \models_4 A(a)$;

   (d) $\langle\Sigma, \Omega\rangle \models_4^c \neg A(a)$ implies $\Sigma \not\models_4 A(a)$.

2. *if $\texttt{CL}(A) \in \Omega$ then for all individuals a,*

   (a) $\Sigma \models_4 A(a)$ implies $\langle\Sigma, \Omega\rangle \models_4^c A(a)$;

   (b) $\Sigma \not\models_4 A(a)$ implies $\langle\Sigma, \Omega\rangle \models_4^c \neg A(a)$.

   Conversely, if $\langle\Sigma, \Omega\rangle$ is satisfiable, then for each primitive concept A,

   (c) $\langle\Sigma, \Omega\rangle \models_4^c A(a)$ implies $\Sigma \models_4 A(a)$;

   (d) $\langle\Sigma, \Omega\rangle \models_4^c \neg A(a)$ implies $\Sigma \not\models_4 A(a)$. ∎

In fact, part 1a of the last Theorem is a special case of Theorem 5 and it has been stated in this form only for symmetry.

Theorem 6 gives us the possibility of comparing our model with Naive CWA, historically the first notion of CWA to be proposed. Naive CWA is defined for finite sets of first-order sentences without equality and whose prenex normal forms contain no existential quantifiers. If T is one such sets, then the naive closure of T, *NCWA(T)*, is given by [34]:

$$\mathit{NCWA(T)} = T \cup \{\neg A : T \not\models A \text{ and } A \in \mathit{HB}(T)\},$$

where $\mathit{HB}(T)$ is the Herbrand Base of T. Now, the first-order translation of a set of Mirlog assertions yields a set of sentences which may contain existential quantification. If we apply the NCWA operator to this kind of theories (and we do not skolemize), the last Theorem tells us that c-entailment on completely closed KBs (*i.e.* all individuals appearing in the KB are closed or all primitives appearing in the KB are closed) is equivalent to Naive CWA for the corresponding first-order theories (of course, negation is applied only to concepts since negated roles are not allowed in the language).

It is worth noting that there is a big methodological difference between our approach and NCWA, or, for that matter, all other approaches with the same goal, as for example in Datalog[1]: in Mirlog, CWA is not something happening *behind the scene*, but is explicitly called upon, via closures, by the document indexer, who has therefore full control of the situation, and is free to apply CWA only on specified terms.

Finally, the reader interested in the relationship between $\models_4, \models, \models_4^c$ from one hand and, from the other hand, the inference relation captured by applying closures to classical KBs, may refer to [37].

4.5 Inconsistencies induced by closures

Let us consider the KB $\langle \Sigma, \Omega \rangle$ where:

$$\begin{aligned} \Sigma &= \{(\mathtt{C} \sqcup \mathtt{D})[\mathtt{a}]\} \\ \Omega &= \{\mathtt{CL}(\mathtt{a})\}. \end{aligned}$$

From an intuitive point of view, the above KB is clearly inconsistent: from one hand, its Σ component asserts that a is either a C or a D without saying which; from the other hand, the Ω component asserts that the knowledge about a is complete, what evidently contradicts Σ's content. As a matter of fact, this KB is also inconsistent from a formal point of view, *i.e.* it has no models. Indeed, let us suppose, to the contrary, that $\mathcal{I}$ is a model of $\langle \Sigma, \Omega \rangle$. As such, $\mathcal{I}$ must satisfy all assertions in Σ, which means that either $t \in \mathtt{C}^{\mathcal{I}}(\mathtt{a}^{\mathcal{I}})$ or $t \in \mathtt{D}^{\mathcal{I}}(\mathtt{a}^{\mathcal{I}})$. Suppose the former is the case. Now, since a is closed, it follows that in every model $\mathcal{J}$ of $\langle \Sigma, \Omega \rangle$, $t \in \mathtt{C}^{\mathcal{J}}(\mathtt{a}^{\mathcal{J}})$. But this is clearly impossible, because C[a] does not follow from the KB.

From a KB with no models, every assertion vacuously follows. This means that closures introduce intolerance to contradictions, a problem that the relevance semantics of Mirlog was designed to solve. This is the price that the model pays for the capability of doing closed-world reasoning. Since it is restricted to certain elements of a KB, we believe it is affordable: it just imposes careful consideration when specifying closures.

4.6 Conclusions

In summary, Mirlog is a description logic with an implication relation that can be broadly characterized by two features:

- first, it does not allow puzzle-mode reasoning, which is not what an IR system is expected to do, thereby gaining capture of relevance and tolerance to inconsistency, which is what an IR system is supposed to need;
- second, it allows selective closed-world reasoning, an important inference mechanism for IR, as it captures the proper way of handling knowledge about document structure and profile.

As such, Mirlog can be seen as an adjustment of a brand of mathematical logics toward the task of IR.

5 Modelling uncertainty

The logic we have described so far is still insufficient for describing *real* retrieval situations, as retrieval is usually not a yes-no question: the representations of documents and queries which the system (and the logic) have access to are inherently imperfect, and the relevance of a document to a query can thus be established only up to a limited degree of certainty. To this end, we extend Mirlog with *fuzzy assertions.*

Fuzzy assertions take inspiration from Zadeh's work on fuzzy sets [58]. A fuzzy set A with respect to a set X is characterized by a membership function $\mu_A : X \rightarrow [0,1]$, assigning an A-membership degree, $\mu_A(x)$, to each element x in X. This membership degree gives us an estimation of the belonging of x to A. Typically, if $\mu_A(x) = 1$ then x definitely belongs to A, while $\mu_A(x) = 0.8$ means that x is "likely" to be an element of A. Moreover, according to Zadeh, the membership function has to satisfy three well-known restrictions, for all $x \in X$ and for all fuzzy sets A, B with respect to X:

$$\begin{aligned} \mu_{A\cap B}(x) &= \min\{\mu_A(x), \mu_B(x)\}, \\ \mu_{A\cup B}(x) &= \max\{\mu_A(x), \mu_B(x)\}, \text{ and} \\ \mu_{\overline{A}}(x) &= 1 - \mu_A(x), \end{aligned}$$

where $\overline{A}$ is the complement of A in X. Other membership functions have been proposed, but it is not our aim to investigate them here (the interested reader can consult *e.g.* [16]).

When we switch to logic, and to DLs in particular, we have terms rather than sets and speak about *degrees of truth* instead of membership degrees. For instance, the assertion that individual `a` is an instance of concept `C`, formally written as `C[a]`, may have as a degree of truth any real number in between 0 and 1: if the degree of truth of `C[a]` is 1, then `a` is definitely an instance of `C`, while if the degree of truth of `C[a]` is 0.8 then `a` is likely to be an instance of `C`. Similarly for role assertions. Hence, in a fuzzy DL, terms become *imprecise* (or *vague*). As a consequence, given a query concept `Q`, the retrieval process produces a ranking of individuals: the rank of `a`, for each individual `a`, is the degree of truth of `Q[a]`, and will be interpreted as the degree of relevance of the document identified by `a` to the query `Q`.

The choice of fuzzy set theory as a way of endowing a DL with the capability to deal with uncertainty is not uncommon [26, 28, 30, 49] and can be motivated both from the syntactical and the semantical point of view. From a semantical point of view, fuzzy logics capture the notion of vague concept, that is a concept that is intrinsically *imprecise* and for which a clear and precise definition is not possible. For instance, "hot" and "tall" are vague concepts. The key fact about vague concepts is that while they are not well defined, assertions involving them may be quite well defined. For instance, the boundaries of the Mount Everest are ill-defined, whereas the assertion stating that the Mount Everest is the highest mountain of the world is clearly definite, and its definiteness is not compromised by the ill-definiteness of the exact boundaries of the mountain. It is easy to see that fuzzy assertions play a key role in content descriptions of documents.

From a proof theoretical point of view, there exist well-known techniques for reasoning in fuzzy logics (see *e.g.* [10, 29, 31]). This is not the case for alternative logics, such as, for instance, probabilistic logics [26, 30, 49]. In particular, [25] shows that probabilistic reasoning is computationally more difficult than non-probabilistic reasoning, and in most cases a complete axiomatization is missing.

Fuzzy logic is not appropriate to deal with *uncertain assertions*, that is assertions which are only true or false, but, due to the lack of precision of the available information, one can only estimate to what extent it is possible or necessary that they are true. For instance, "line", and "polygon" are precise concepts, but due to the lack of precision of the available information we may only be able to estimate to what degree a certain object in an image is *e.g.* a polygon. The logics dealing with this kind of uncertainty have been called *Possibilistic Logics* [17]. Possibilistic DLs are discussed in [28].

The combination of possibilistic and fuzzy logic would lead to the treatment of *uncertain fuzzy assertions*, *i.e.* fuzzy assertions for which the available reference information is not precise. While this combination is possible, and maybe even desirable for IR purposes, our model only provides fuzzy assertions. A DL allowing uncertain fuzzy assertions can be obtained by combining the approach in [28] with fuzzy Mirlog.

For a better readability, we will first give syntax and 2-valued semantics of MIRLOG with no closures. This step is rather straightforward as we will use the most popular and classic version of fuzzy logic based on the $\min - \max$ functions introduced by Zadeh. In particular, our logic is a DL version of the formal framework described in [10, 29, 31, 57]. In two successive steps, we extend the resulting logic to the 4-valued semantics and to closures.

5.1 Syntax and 2-valued fuzzy semantics

A *fuzzy assertion* is an expression of type $\langle\alpha \geq n\rangle$ or of type $\langle\alpha > n\rangle$, where α is an assertional formula or a definition (as defined in Section 2) and $n \in [0, 1]$. We will confine ourselves with the former kind of assertions, as the extension to the latter is trivial.

The intended meaning of *e.g.* $\langle\alpha \geq n\rangle$ is "the degree of truth of α is *at least* n". For instance, $\langle\texttt{Tall}[\texttt{umberto}] \geq 0.7\rangle$ means that the degree of truth of $\texttt{Tall}[\texttt{umberto}]$ is at least 0.7 (*i.e.* $\texttt{umberto}$ is likely to be tall). Formally, an interpretation is a triple $\mathcal{I} = (\Delta^{\mathcal{I}}, (\cdot)^{\mathcal{I}}, |\cdot|_{\mathcal{I}})$, where:

1. $\Delta^{\mathcal{I}}$, the domain of $\mathcal{I}$, is a non-empty set;
2. $(\cdot)^{\mathcal{I}}$, the interpretation function of $\mathcal{I}$, maps each fuzzy assertion into $\{t, f\}$;
3. $|\cdot|_{\mathcal{I}}$, the *fuzzy valuation*, maps each concept into a function from $\Delta^{\mathcal{I}}$ into $[0, 1]$, and each role into a function from $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ into $[0, 1]$ (for a concept C, $|C|_{\mathcal{I}}$ can be seen as the membership degree function of C),

such that:

$$\begin{array}{rcl}
|\top|_{\mathcal{I}}(d) & = & 1 \\
|\perp|_{\mathcal{I}}(d) & = & 0 \\
|C_1 \sqcap C_2|_{\mathcal{I}}(d) & = & \min\{|C_1|_{\mathcal{I}}(d), |C_2|_{\mathcal{I}}(d)\} \\
|C_1 \sqcup C_2|_{\mathcal{I}}(d) & = & \max\{|C_1|_{\mathcal{I}}(d), |C_2|_{\mathcal{I}}(d)\} \\
|\neg C|_{\mathcal{I}}(d) & = & 1 - |C|_{\mathcal{I}}(d) \\
|\forall R.C|_{\mathcal{I}}(d) & = & \min_{d' \in \Delta^{\mathcal{I}}}\{\max\{1 - |R|_{\mathcal{I}}(d, d'), |C|_{\mathcal{I}}(d')\}\} \\
|\exists R.C|_{\mathcal{I}}(d) & = & \max_{d' \in \Delta^{\mathcal{I}}}\{\min\{|R|_{\mathcal{I}}(d, d'), |C|_{\mathcal{I}}(d')\}\}
\end{array}$$

and:

$$\begin{array}{rcl}
\langle C[a] \geq n\rangle^{\mathcal{I}} = t & \text{iff} & |C|_{\mathcal{I}}(a^{\mathcal{I}}) \geq n \\
\langle R[a, b] \geq n\rangle^{\mathcal{I}} = t & \text{iff} & |R|_{\mathcal{I}}(a^{\mathcal{I}}, b^{\mathcal{I}}) \geq n \\
\langle C_1 \sqsubseteq C_2 \geq n\rangle^{\mathcal{I}} = t & \text{iff} & \min_{d \in \Delta^{\mathcal{I}}}\{|\neg C_1 \sqcup C_2|_{\mathcal{I}}(d)\} \geq n.
\end{array}$$

As far as the constraints on the fuzzy valuation, they are standard for conjunction, disjunction, and negation. The constraint on the universal quantifier is the result of viewing $\forall R.C$ as the open first order formula $\forall y(R(x, y) \Rightarrow C(y))$ or, equivalently, as $\forall y(\neg R(x, y) \vee C(y))$. Now, in classical logic a formula like $\forall x A$ is interpreted as a conjunction over the elements of the domain of the interpretation. Then, the derivation of the constraint for $\forall R.C$ is just a matter of applying the constraints defined for $\sqcap$, $\sqcup$ and $\neg$ to the classical view of universal quantification. As a quality assessment of the resulting constraint, observe that the semantics of the $\forall$ operator is such that $|\forall R.C|_{\mathcal{I}}(d) \geq n$ iff for all $d' \in \Delta^{\mathcal{I}}$ if $|R|_{\mathcal{I}}(d, d') > 1 - n)$ then $|C|_{\mathcal{I}}(d') \geq n$. Analogously, $\exists R.C$ is viewed as $\exists y(R(x, y) \wedge C(y))$, and $\exists y A$ as disjunction over the elements in the domain of the interpretation.

As far as the interpretation function is concerned, the semantics of the assertion $\langle C_1 \sqsubseteq C_2 \geq n\rangle$ is a consequence of viewing the definition $C_1 \sqsubseteq C_2$ as the implication $\forall x(C_1(x) \Rightarrow C_2(x))$.

Fuzzy satisfiability, fuzzy equivalence and fuzzy entailment are defined as natural extensions of the corresponding non-fuzzy notions. In particular, a fuzzy interpretation $\mathcal{I}$ *satisfies* (*is a model of*) a fuzzy assertion $\langle\alpha \geq n\rangle$ iff $\langle\alpha \geq n\rangle^{\mathcal{I}} = t$. $\mathcal{I}$ *satisfies* (*is a model of*) a set of fuzzy assertions (a *fuzzy KB*) Σ iff it satisfies all assertions in Σ.

A fuzzy KB Σ *entails* a fuzzy assertion $\langle\alpha \geq n\rangle$ (written $\Sigma \models^f \langle\alpha \geq n\rangle$) iff all models of Σ satisfy $\langle\alpha \geq n\rangle$. Given a fuzzy KB Σ and a crisp (*i.e.* non-fuzzy) assertion α, we define the *maximal degree of truth* of α with respect to Σ (written $Maxdeg(\Sigma, \alpha)$) to be the maximal $n \geq 0$ such that $\Sigma \models^f \langle\alpha \geq n\rangle$.

A very important property of the semantics defined so far is stated by the following proposition, which is a straightforward application of Lee's work [31] to the DL case.

Proposition 7 *Let Σ be a set of fuzzy assertions of type $\langle\alpha \geq n\rangle$, where $n > 0.5$. Let $\overline{\Sigma}$ be $\{\alpha : \langle\alpha \geq n\rangle \in \Sigma\}$. Then there is an $m > 0.5$ such that $\Sigma \models^f \langle\beta \geq m\rangle$ iff $\overline{\Sigma} \models \beta$.*

It can be verified that the above proposition does not hold if some n appearing in Σ is ≤ 0.5. For instance,

$$\{\langle \mathtt{A[a]} \geq 0.3\rangle, \langle(\neg\mathtt{A} \sqcup \mathtt{B})\mathtt{[a]} \geq 0.6\rangle\} \not\models^f \langle \mathtt{B[a]} \geq n\rangle$$

for all $n > 0$, whereas

$$\{\mathtt{A[a]}, (\neg\mathtt{A} \sqcup \mathtt{B})\mathtt{[a]}\} \models \mathtt{B[a]}.$$

In the following we will assume that the values n occurring in a Mirlog KB are greater than 0.5. This is not a limitation as each value n can be normalized by means of the formula $n := \frac{n+1}{2}$.

5.2 Relevance fuzzy semantics

Consistently with our approach of distinguishing explicit from implicit falsehood (*e.g.* distinguishing $f \in C^{\mathcal{I}}(a^{\mathcal{I}})$ from $t \not\in C^{\mathcal{I}}(a^{\mathcal{I}})$), the relevance, 4-valued semantics of Mirlog is based on two fuzzy valuations: $|\cdot|^+_{\mathcal{I}}$ and $|\cdot|^-_{\mathcal{I}}$. $|C|^+_{\mathcal{I}}(a^{\mathcal{I}})$ will be interpreted as the *degree of truth* of $C[a]$, whereas $|C|^-_{\mathcal{I}}(a^{\mathcal{I}})$ will analogously be interpreted as the *degree of falsity* of $C[a]$. As we have seen, in classical "two-valued" fuzzy systems: $|\cdot|^-_{\mathcal{I}} = 1 - |\cdot|^+_{\mathcal{I}}$. In the 4-valued case, instead, we may well have $|C|^+_{\mathcal{I}}(d) = 0.6$ and $|C|^-_{\mathcal{I}}(d) = 0.8$. This is a natural consequence of our 4-valued approach.

Formally, a 4-valued interpretation is a 4-tuple $\mathcal{I} = (\Delta^{\mathcal{I}}, (\cdot)^{\mathcal{I}}, |\cdot|^+_{\mathcal{I}}, |\cdot|^-_{\mathcal{I}})$, where:

1. $\Delta^{\mathcal{I}}$, the domain of $\mathcal{I}$, is a non-empty set;

2. $(\cdot)^{\mathcal{I}}$, the interpretation function of $\mathcal{I}$, maps each fuzzy assertion into $\{t, f\}$;

3. $|\cdot|^+_{\mathcal{I}}$, the *positive fuzzy valuation*, maps each concept into a function from $\Delta^{\mathcal{I}}$ into $[0, 1]$, and each role into a function from $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ into $[0, 1]$; and

4. $|\cdot|^-_{\mathcal{I}}$, the *negative fuzzy valuation*, maps each concept into a function from $\Delta^{\mathcal{I}}$ into $[0, 1]$, and each role into a function from $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ into $[0, 1]$

such that:

$$\begin{array}{rcl}
|\top|^+_{\mathcal{I}}(d) & = & 1\\
|\top|^-_{\mathcal{I}}(d) & = & 0\\
|\perp|^+_{\mathcal{I}}(d) & = & 0\\
|\perp|^-_{\mathcal{I}}(d) & = & 1\\
|C_1 \sqcap C_2|^+_{\mathcal{I}}(d) & = & \min\{|C_1|^+_{\mathcal{I}}(d), |C_2|^+_{\mathcal{I}}(d)\}\\
|C_1 \sqcap C_2|^-_{\mathcal{I}}(d) & = & \max\{|C_1|^+_{\mathcal{I}}(d), |C_2|^+_{\mathcal{I}}(d)\}\\
|C_1 \sqcup C_2|^+_{\mathcal{I}}(d) & = & \max\{|C_1|^+_{\mathcal{I}}(d), |C_2|^+_{\mathcal{I}}(d)\}\\
|C_1 \sqcup C_2|^-_{\mathcal{I}}(d) & = & \min\{|C_1|^+_{\mathcal{I}}(d), |C_2|^+_{\mathcal{I}}(d)\}\\
|\neg C|^+_{\mathcal{I}}(d) & = & |C|^-_{\mathcal{I}}(d)\\
|\neg C|^-_{\mathcal{I}}(d) & = & |C|^+_{\mathcal{I}}(d)\\
|\forall R.C|^+_{\mathcal{I}}(d) & = & \min_{d' \in \Delta^{\mathcal{I}}}\{\max\{1 - |R|^+_{\mathcal{I}}(d,d'), |C|^+_{\mathcal{I}}(d')\}\}\\
|\forall R.C|^-_{\mathcal{I}}(d) & = & \max_{d' \in \Delta^{\mathcal{I}}}\{\min\{|R|^+_{\mathcal{I}}(d,d'), |C|^-_{\mathcal{I}}(d')\}\}\\
|\exists R.C|^+_{\mathcal{I}}(d) & = & \max_{d' \in \Delta^{\mathcal{I}}}\{\min\{|R|^+_{\mathcal{I}}(d,d'), |C|^+_{\mathcal{I}}(d')\}\}\\
|\exists R.C|^-_{\mathcal{I}}(d) & = & \min_{d' \in \Delta^{\mathcal{I}}}\{\max\{1 - |R|^+_{\mathcal{I}}(d,d'), |C|^-_{\mathcal{I}}(d')\}\}
\end{array}$$

and

$$\begin{array}{rcl}
t \in \langle C[a] \geq n\rangle^{\mathcal{I}} & \text{iff} & |C|^+_{\mathcal{I}}(a^{\mathcal{I}}) \geq n\\
f \in \langle C[a] \geq n\rangle^{\mathcal{I}} & \text{iff} & |C|^-_{\mathcal{I}}(a^{\mathcal{I}}) \geq n\\
t \in \langle R[a,b] \geq n\rangle^{\mathcal{I}} & \text{iff} & |R|_{\mathcal{I}}(a^{\mathcal{I}}, b^{\mathcal{I}}) \geq n\\
f \in \langle R[a,b] \geq n\rangle^{\mathcal{I}} & \text{iff} & |R|^-_{\mathcal{I}}(a^{\mathcal{I}}, b^{\mathcal{I}}) \geq n\\
t \in \langle C_1 \sqsubseteq C_2 \geq n\rangle^{\mathcal{I}} & \text{iff} & \min_{d \in \Delta^{\mathcal{I}}}\{\max\{1 - |C_1|^+_{\mathcal{I}}(d), |C_2|^+_{\mathcal{I}}(d)\}\} \geq n\\
f \in \langle C_1 \sqsubseteq C_2 \geq n\rangle^{\mathcal{I}} & \text{iff} & \max_{d \in \Delta^{\mathcal{I}}}\{\min\{|C_1|^+_{\mathcal{I}}(d), |C_2|^-_{\mathcal{I}}(d)\}\} \geq n
\end{array}$$

The semantics for the $\forall$ and $\exists$ operators are such that $|\forall R.C|^+_{\mathcal{I}} = |\exists R.\neg C|^-_{\mathcal{I}}$ and $|\exists R.C|^+_{\mathcal{I}} = |\forall R.\neg C|^-_{\mathcal{I}}$. Moreover, the semantics reflects the definition of the two-valued case. Hence, we allow *modus ponens* on roles. A similar argument holds for $\langle C_1 \sqsubseteq C_2 \geq n\rangle$.

For brevity, we do not state the notions of satisfaction and 4-valued entailment ($\models^f_4$); they are the obvious translation of the corresponding notions introduced in the previous section.

As an example, let us consider a KB about two images $\mathtt{i}$ and $\mathtt{j}$ whose content is described by means of the following assertions and (background) definitions:

$$\begin{array}{c}
\langle \mathtt{About}[\mathtt{i},\mathtt{a}] \geq 0.8\rangle,\ \langle \mathtt{DonGiovanni}[\mathtt{a}] \geq 1\rangle,\\
\langle \mathtt{About}[\mathtt{j},\mathtt{b}] \geq 0.7\rangle,\ \langle \mathtt{WestSideStory}[\mathtt{b}] \geq 1\rangle\\
\langle \mathtt{DonGiovanni} \sqsubseteq \mathtt{EuropeanOpera} \geq 1\rangle,\ \langle \mathtt{WestSideStory} \sqsubseteq \mathtt{AmericanOpera} \geq 1\rangle,\\
\langle \mathtt{EuropeanOpera} \sqsubseteq \mathtt{Opera} \sqcap (\exists \mathtt{ConductedBy.European}) \geq 0.9\rangle,\\
\langle \mathtt{AmericanOpera} \sqsubseteq \mathtt{Opera} \sqcap (\exists \mathtt{ConductedBy.European}) \geq 0.8\rangle.
\end{array}$$

Suppose a user is interested in retrieving those images that are about an opera conducted by a European director. To this end, the query:

$$\exists \mathtt{About}.(\mathtt{Opera} \sqcap \exists \mathtt{ConductedBy.European})$$

can be used. It can be verified that the maximal degree of truth attributed to $\mathtt{i}$ is 0.8, whereas that of $\mathtt{j}$ is 0.7.

Analogously to the two-valued case, the following Proposition holds. It allows to import in the present context, the properties of 4-valued semantics discussed in Section 3.

Proposition 8 *Let Σ be a set of fuzzy assertions of type $\langle \alpha \geq n\rangle$, where $n > 0.5$. Let $\overline{\Sigma}$ be $\{\alpha : \langle \alpha \geq n\rangle \in \Sigma\}$. Then there is a $m > 0.5$ such that $\Sigma \models^f_4 \langle \beta \geq m\rangle$ iff $\overline{\Sigma} \models_4 \beta$.*

5.3 Extension to closures

The treatment of closures is straightforward. We will give only a brief description without going into the details, as they are tedious and can easily be worked out.

Satisfiability of closures is defined on the basis of *fuzzy epistemic interpretations.* Formally, a fuzzy epistemic interpretation is a pair $\langle \mathcal{I}, \mathcal{W} \rangle$, where $\mathcal{I}$ is a fuzzy interpretation and $\mathcal{W}$ is a set of fuzzy interpretations defined on the same domain Δ and mapping, as for the crisp case, the same individuals to the same objects.

Definition 6 *An* epistemic interpretation $\langle \mathcal{I}, \mathcal{W} \rangle$ satisfies *a primitive closure* $\mathtt{CL}(a)$ *if and only if the following conditions hold for all* $n \geq 0$*:*

1. *for every primitive concept symbol A, $|A|^+_{\mathcal{I}}(\gamma(a)) \geq n$ iff $|A|^+_{\mathcal{J}}(\gamma(a)) \geq n$ for all $\mathcal{J} \in \mathcal{W}$;*
2. *for every primitive concept symbol A, $|A|^-_{\mathcal{I}}(\gamma(a)) \geq n$ iff $|A|^+_{\mathcal{J}}(\gamma(a)) < n$ for some $\mathcal{J} \in \mathcal{W}$;*
3. *for every primitive role symbol P and parameter $p \in \Delta$, $|P|^+_{\mathcal{I}}(\gamma(a), p) \geq n$ iff $|P|^+_{\mathcal{J}}(\gamma(a), p) \geq n$ for all $\mathcal{J} \in \mathcal{W}$;*
4. *for every primitive role symbol P and parameter $p \in \Delta$, $|P|^-_{\mathcal{I}}(\gamma(a), p) \geq n$ iff $|P|^+_{\mathcal{J}}(\gamma(a), p) < n$ for some $\mathcal{J} \in \mathcal{W}$.*

A fuzzy epistemic interpretation satisfies *(is a* model *of) a set of closures if and only if it satisfies each closure in the set.* ∎

Finally, satisfiability of a fuzzy KB $\langle \Sigma, \Omega \rangle$ and fuzzy c-entailment ($\models^{cf}_4$) are defined as for the crisp case. It is easy to verify that, for any model $\mathcal{I}$ of a KB $\langle \Sigma, \Omega \rangle$ and closed individual a, $\gamma(a)$ is such that $|A|^+_{\mathcal{I}}(\gamma(a)) \geq n$ just in case $\langle A[a] \geq n \rangle$ is entailed by Σ, in symbols $\Sigma \models^f_4 \langle A[a] \geq n \rangle$.

It follows that fuzzy c-entailment exhibits similar properties to those of crisp c-entailment. For instance, the fuzzy version of Proposition 2 is as follows:

Proposition 9 *Let $\langle \Sigma, \Omega \rangle$ be a KB,* $\mathtt{CL}$*(a)*$\in \Omega$. *Then*

1. *either $\langle \Sigma, \Omega \rangle \models^{cf}_4 \langle C[a] \geq n \rangle$ or $\langle \Sigma, \Omega \rangle \models^{cf}_4 \langle \neg C[a] \geq n \rangle$, for any quantifier free C;*
2. *if $\langle \Sigma, \Omega \rangle$ is completely closed, then either $\langle \Sigma, \Omega \rangle \models^{cf}_4 \langle C[a] \geq n \rangle$ or $\langle \Sigma, \Omega \rangle \models^{cf}_4 \langle \neg C[a] \geq n \rangle$, for any C.* ∎

6 Reasoning in Mirlog

The decision problems considered important in 2-valued DLs, notably the instance checking and subsumption problem, can be reduced to the KB satisfiability problem. In fact, it is easily verified that:

$$C \sqsubseteq D \text{ iff } \{C(a)\} \models D(a)$$

for any individual a not occurring in $C \sqcup D$, and

$$\Sigma \models C(a) \text{ iff } \Sigma \cup \{(\neg C)(a)\} \text{ is not satisfiable} \tag{12}$$

There exists a well known sound and complete algorithm based on constraint propagation [47], which is essentially an analytic tableaux-based decision procedure, for deciding KB satisfiability. This proof method has also allowed the derivation of many complexity results concerning 2-valued DLs (see *e.g.* [8, 13])[11].

If we switch to a 4-valued setting, we need an alternative proof procedure as relation (12) no longer holds. There exists a well known subsumption testing procedure, which is a DL adaption of Levesque's algorithm [32] for entailment [6, 40, 43]. The algorithm performs structural subsumption in a efficient way, but has several drawbacks. First, it does not work within our semantics. Second, it cannot be used for the instance checking test as this problem is in a higher complexity class than the subsumption problem. Third, it is rather difficult to adapt the algorithm to a DL with an even slightly different set of term-forming operators; in general, this is the price we must pay if we want fast special purpose algorithms.

For these reasons, we have developed a sequent calculus-based proof procedure for instance checking that solves the subsumption problem too [38]. With a minor modification, this calculus can be used to test subsumption in 2-valued semantics, in which case it shows the same performance as the above mentioned structural subsumption algorithm. Moreover, the method is easily adaptable to the different DLs described in the literature. For space reason, we do not present this proof procedure here, the interested reader may refer to [37].

On the basis of this method, it has been proven [51] that deciding entailment ($\models_4$) for a language with closures but without definitions is a PSPACE-complete problem, while the same problem becomes EXPTIME-hard when definitions are considered.

Recently, it has been shown that analytic tableaux methods for two-valued DLs are quite inefficient, as the length of the proof of a formula may be exponential in the length of the formula rather than in the number of different letters occurring in it [23, 22]. In fact, consider the formula:

$$\alpha = (A \vee B) \wedge (\sim A \vee B) \wedge (\sim A \vee \sim B).$$

α has 2 different letters, which means that it has 4 possible two-valued interpretations. As a consequence, a semantic based decision procedure like the Davis-Longemann-Loveland [12], can test its satisfiability after enumerating at most 4 interpretations. On the other hand, an analytic tableaux calculus [20] using the two rules:

$$\wedge\text{-rule}\ \frac{A, B}{A \wedge B} \qquad\qquad \vee\text{-rule}\ \frac{A \quad B}{A \vee B} \tag{13}$$

generates a proof tree with $O(2^3)$ leafs, as shown in Figure 3. Essentially, each path from a leaf to the root of the tree is an attempt to build a model of the formula α. The paths marked with a $\times$ are failed attempts, as they contain both a propositional letter and its negation. Each path marked with a w_i represents a model of the formula: in Figure 3 there are two such paths, marked w_1 and w_2, sanctioning the satisfiability of α.

In deciding the satisfiability of a formula, an analytic tableaux method performs *syntactic branching*, that is, a branching guided by the syntactic structure of the formula under consideration. As discussed in [11], any application of the $\vee - rule$ may generate two subtrees which are *not mutually inconsistent*, that is two subtrees which may share models. This is the case of the subtrees generated from the node marked with a $*$ in Figure 3, which both generate the only model of α, given by $\{\sim A, B\}$. So, the set of interpretations enumerated by analytic tableaux procedures are intrinsically redundant. As a consequence, the number of interpretations generated grows exponentially with the number of disjunctions occurring in the formula to be proven, although the number of different interpretations is much smaller. This redundancy is a source of inefficiency. Unfortunately, this inefficiency

[11] An exhaustive list of results can be found in the DL WWW home page at `http://www.dl.kr.org/dl`.

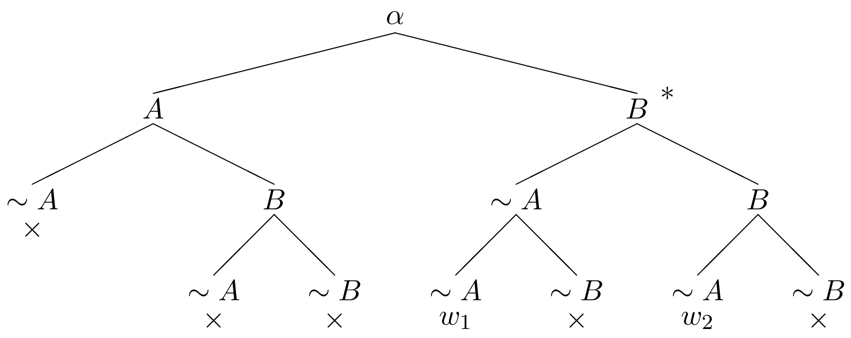

Figure 3: Tableaux for $\alpha = (A \vee B) \wedge (\sim A \vee B) \wedge (\sim A \vee \sim B)$.

carries over our sequent calculus, which is just a generalization of the analytic tableaux calculus for DLs.

In order to solve this problem, semantic-based methods have been proposed for two-valued DLs [23], inspired by a calculus, named **KE** [11], that does not exhibit the over-generation problem discussed above. We are currently developing a semantic-based calculus for Mirlog.

7 Conclusions

We have presented a description logic tailored on the requirements of information retrieval modelling. In particular, the logic is based on relevance semantics and allows a form of non-monotonic reasoning. It has been argued at length how these features match with the task of information retrieval, thus providing ground for their inclusion in Mirlog.

The computational aspects of Mirlog have been also discussed, even though for reasons of space, we could not present the sequent calculus that we have developed for reasoning on Mirlog knowledge bases. As it has been pointed out, a more efficient calculus is being designed and implemented, based on recent insights on proof theory for description logic.

The driving motivation of our work has been the realization of a model of information retrieval that goes beyond the still prevailing keyword-based approach. The work reported in this paper makes only a first step towards this end, although a necessary and non-trivial one. Namely, it proposes itself as a *tool* for performing the retrieval of information in a way that departs from traditional approaches. In order to carry on the ambitious program that we have set up, at least two more steps are needed.

The first step, concerns the *usage* of the Mirlog tool. This means that a model of information retrieval has to be defined, which specifies in rigorous terms how Mirlog must be employed in representing documents. Indeed, the notion of Mirlog concept is a sufficiently precise specification of how a user query is to be expressed, and the implication relation of Mirlog is a sufficiently precise specification of the conditions under which a document ought to be retrieved. However, all we have said about the representation of a document is that it is a set of assertions, and this is clearly still a too vague notion for putting Mirlog at work in a realistic setting. We have started to work on this subject as far as image documents are concerned. Preliminary results may be found in [35].

A second, important step, regards the *evaluation* of the resulting model. In developing Mirlog, we have given paramount importance to the computational aspect of the problem. More specifically, we have oriented our choice towards a description logic *also* because of the basic decidability results that were known for these logics. In addition, we have studied the complexity of the decision problem of Mirlog, being able to prove that, from one hand, the logic is decidable, while from the other, it has exponential worst case complexity. This latter result does not seem to be particularly encouraging. However, it is important to realize that any significantly expressive formalism is plagued by results of this sort, if not worse. A practical evaluation of Mirlog is needed, in order to gain a finer understanding of the computational behaviour of the logic in the "average" IR application and, at the same time, to observe the effectiveness of Mirlog syntax and semantics in coping with document representation and retrieval. This field is, at the moment, totally unexplored: to the best of our knowledge there has been no systematic attempt to use logic for modelling documents and their contents in the way we plan.

A successful evaluation would then raise the problem of developing a *methodology* for applying a Mirlog-based model to a realistic problem. Description logics have been employed in various applications where a knowledge representation service was required. As a result, there have been studies on the definition of a methodology for engineering knowledge in the form of a DL knowledge base (a general methodology for knowledge based systems is presented in [7], while [5] makes the case for data management). However, the specificity of information retrieval requires a refined methodology that tackle issues such as: what level of granularity should the lexical knowledge employed in retrieval have, where to find it, how to input it into a system; or, how does one deal with document standards such as SGML, HTML and the like; or, how to import and use knowledge about document layout, also pervaded by standards. And so on. Most of these problems have simple solutions, yet a standard way of approaching them is needed, possibly supported by automatic tools.

Both the practical evaluation and the methodology development require a tight interaction with the world of library and information science. We hope that the advent of digital libraries will put this world in a closer contact with our own.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of databases.* Addison-Wesley, New York, NY, 1995.

[2] Alan R. Anderson and Nuel D. Belnap. *Entailment - the logic of relevance and necessity*, volume 1. Princeton University Press, Princeton, NJ, 1975.

[3] Nicholas J. Belkin. Ineffable concepts in information retrieval. In Karen Sparck Jones, editor, *Information retrieval experiment*, pages 44–58. Butterworths, London, UK, 1981.

[4] Nuel D. Belnap. How a computer should think. In Gilbert Ryle, editor, *Contemporary aspects of philosophy*, pages 30–56. Oriel Press, Stocksfield, UK, 1977.

[5] Alexander Borgida. Description logics in data management. *IEEE Transactions on Data and Knowledge Engineering*, 7(5):671–682, 1995.

[6] Alexander Borgida and Peter F. Patel-Schneider. A semantics and complete algorithm for subsumption in the CLASSIC description logic. *Journal of Artificial Intelligence Research*, 1:277–308, 1994.

[7] Ronal J. Brachman, Deborah L. McGuiness, Peter F. Patel-Schneider, Lori Alperin Resnick, and Alexander Borgida. Living with CLASSIC: when and how to use a KL-ONE-like language. In John Sowa, editor, *Principles of Semantic Networks.* Morgan Kaufmann, 1990.

[8] Martin Buchheit, Francesco M. Donini, and Andrea Schaerf. Decidable reasoning in terminological knowledge representation systems. *Journal of Artificial Intelligence Research*, 1:109–138, 1993.

[9] Paolo Buongarzoni, Carlo Meghini, Rossella Salis, Fabrizio Sebastiani, and Umberto Straccia. Logical and computational properties of the description logic MIRTL. In Alexander Borgida, Maurizio Lenzerini, Daniele Nardi, and Bernhard Nebel, editors, *Proceedings of DL-95, 4th International Workshop on Description Logics*, pages 80–84, Roma, Italy, 1995.

[10] Jianhua Chen and Sukhamany Kundu. A sound and complete fuzzy logic system using Zadeh's implication operator. In Zbigniew W. Ras and Michalewicz Maciek, editors, *Proc. of the 9th Int. Sym. on Methodologies for Intelligent Systems (ISMIS-96)*, number 1079 in Lecture Notes In Artificial Intelligence, pages 233–242. Springer-Verlag, 1996.

[11] Marcello D'Agostino and Marco Mondadori. The taming of the cut. Classical refutations with analytical cut. *Journal of Logic and Computation*, 4(3):285–319, 1994.

[12] M. Davis, G. Longemann, and D. Loveland. A machine program for theorem proving. *Journal of the ACM*, 5(7):394–397, 1962.

[13] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, and Werner Nutt. The complexity of concept languages. In *Proceedings of KR-91, 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 151–162, Cambridge, MA, 1991.

[14] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, Werner Nutt, and Andrea Schaerf. Adding epistemic operators to concept languages. In *KR-92*, pages 342–353. Morgan Kaufmann, 1992.

[15] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, Werner Nutt, and Andrea Schaerf. Queries, rules and definitions as epistemic sentences in concept languages. In *Proceedings of the ECAI-94 Workshop on Knowledge Representation and Reasoning*, number 810 in Lecture Notes in Computer Science, pages 113–132. Springer, 1994.

[16] Didier Dubois and Henri Prade. *Fuzzy Sets and Systems.* Academic Press, New York, NJ, 1980.

[17] Didier Dubois and Henri Prade. Possibilistic logic. In Dov M. Gabbay and C. J. Hogger, editors, *Handbook of Logic in Artificial Intelligence*, volume 3, pages 439–513. Clarendon Press, Oxford, Dordrecht, NL, 1986.

[18] J. Michael Dunn. Intuitive semantics for first-degree entailments and coupled trees. *Philosophical Studies*, 29:149–168, 1976.

[19] J. Michael Dunn. Relevance logic and entailment. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, pages 117–224. Reidel, Dordrecht, NL, 1986.

[20] Melvin Fitting. *First-Order Logic and Automated Theorem Proving.* Springer-Verlag, 1990.

[21] Michael Gelfond and Halina Przymusinska. Negation as failure: careful closure procedure. *Artificial Intelligence*, 30:273–287, 1986.

[22] Fausto Giunchiglia and Roberto Sebastiani. Buiding decision procedures for modal logics from propositional decision procedures - the case study of modal K. In *Proc. of the 13th Conf. on Automated Deduction (CADE-96)*, number 449 in Lecture Notes In Artificial Intelligence. Springer-Verlag, 1996.

[23] Franco Giunchiglia and Roberto Sebastiani. A SAT-based decision procedure for ALC. In *Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR-96)*, 1996.

[24] Susan Haack. *Philosophy of logics.* Cambridge University Press, Cambridge, UK, 1978.

[25] Joseph Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

[26] Jochen Heinsohn. Probabilistic description logics. In R. Lopez de Mantara and D. Pool, editors, *Proceedings of the 10th Conference on Uncertainty in Artificila Intelligence*, pages 311–318, 1994.

[27] Jerry R. Hobbs and Stanley J. Rosenschein. Making computational sense of Montague's intensional logic. *Artificial Intelligence*, 9:287–306, 1978.

[28] Bernhard Hollunder. An alternative proof method for possibilistic logic and its application to terminological logics. In *10th Annual Conference on Uncertainity in Artificial Intelligence*, pages – , Seattle, Washington, 1994. R. Lopez de Mantaras and D. Pool.

[29] Mitsuru Ishizuka and Naoki Kanai. Prolog-ELF: incorporating fuzzy logic. In *Proc. of the 9th Int. Joint Conf. on Artificial Intelligence (IJCAI-85)*, pages 701–703, Los Angeles, CA, 1985.

[30] Manfred Jäger. Probabilistic reasoning in terminological logics. In *Proceedings of KR-94, 5-th International Conference on Principles of Knowledge Representation and Reasoning*, pages 305–316, Bonn, FRG, 1994.

[31] Richard C. T. Lee. Fuzzy logic and the resolution principle. *Journal of the ACM*, 19(1):109–119, January 1972.

[32] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of AAAI-84, 4th Conference of the American Association for Artificial Intelligence*, pages 198–202, Austin, TX, 1984.

[33] Hector J. Levesque. Logic and the complexity of reasoning. *Journal of Philosophical Logic*, 17:355–389, 1988.

[34] Witold Łukaszewicz. *Nonmonotonic reasoning: formalization of commonsense reasoning.* Ellis Horwood, Chichester, UK, 1990.

[35] Carlo Meghini, Fabrizio Sebastiani, and Umberto Straccia. The terminological image data model (extended abstract). Technical Report B4-32-11-96, Istituto di Elaborazione della Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy, December 1996.

[36] Carlo Meghini, Fabrizio Sebastiani, Umberto Straccia, and Costantino Thanos. A model of information retrieval based on a terminological logic. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 298–307, Pittsburgh, PA, 1993. Published by ACM Press, Baltimore, MD.

[37] Carlo Meghini and Umberto Straccia. Information retrieval: Foundations of a description logic-based approach. Technical Report B4-18-06-96, Istituto di Elaborazione della Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy, June 1996.

[38] Carlo Meghini and Umberto Straccia. A relevance terminological logic for information retrieval. In *Proceedings of SIGIR-96, 19th International Conference on Research and Development in Information Retrieval*, pages 197–205, Zurich, Switzerland, 1996.

[39] E.J. Nelson. On three logical principles in intension. *The Monist*, 43, 1933.

[40] Peter F. Patel-Schneider. A four-valued semantics for frame-based description languages. In *Proceedings of AAAI-86, 5th Conference of the American Association for Artificial Intelligence*, pages 344–348, Philadelphia, PA, 1986.

[41] Peter F. Patel-Schneider. A hybrid, decidable, logic-based knowledge representation system. *Computational Intelligence*, 3:64–77, 1987.

[42] Peter F. Patel-Schneider. A hybrid, decidable, logic-based knowledge representation system. *Computational Intelligence*, 3:64–77, 1987.

[43] Peter F. Patel-Schneider. A four-valued semantics for terminological logics. *Artificial Intelligence*, 38:319–351, 1989.

[44] Raymond Reiter. On closed-world data bases. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, pages 55–76. Plenum Press, 1978.

[45] Raymond Reiter. On asking what a database knows. In J.W. Lloyd, editor, *Proceedings of the Symposium on Computational Logic*, pages 96–113. Springer Verlag, 1990.

[46] Tefko Saracevic. Relevance: a review of and a framework for thinking on the notion of information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

[47] Manfred Schmidt-Schauß and Gert Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48:1–26, 1991.

[48] Fabrizio Sebastiani. A probabilistic terminological logic for modelling information retrieval. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 122–130, Dublin, IRL, 1994. Published by Springer Verlag, Heidelberg, FRG.

[49] Fabrizio Sebastiani. A probabilistic terminological logic for modelling information retrieval. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 122–130, Dublin, IRL, 1994. Published by Springer Verlag, Heidelberg, FRG.

[50] Umberto Straccia. Document retrieval by relevance terminological logics. In Ian Ruthven, editor, *Proceedings of MIRO-95, Workshop on Multimedia Information Retrieval*, Glasgow, UK, 1996. Springer Verlag, Heidelberg, FRG.

[51] Umberto Straccia. A sequent calculus for reasoning in four-valued description logics. In *Proc. of the Int. Conf. on Analytic Tableaux and Related Methods (TABLEAUX-97)*, Pont-à-Mousson, France, 1997. To appear.

[52] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, GB, second edition, 1979.

[53] Cornelis J. van Rijsbergen. A new theoretical framework for information retrieval. In *Proceedings of SIGIR-86, 9th ACM International Conference on Research and Development in Information Retrieval*, pages 194–200, Pisa, Italy, 1986.

[54] Cornelis J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.

[55] Cornelis J. van Rijsbergen. Towards an information logic. In *Proceedings of SIGIR-89, 12th ACM International Conference on Research and Development in Information Retrieval*, pages 77–86, Cambridge, MA, 1989.

[56] Gerd Wagner. Ex contradictione nihil sequitur. In *Proceedings of IJCAI-91, 12th International Joint Conference on Artificial Intelligence*, pages 538–543, Sidney, Australia, 1991.

[57] Ronald R. Yager. Fuzzy sets as a tool for modeling. In Jan van Leeuwen, editor, *Computer Science Today*, number 1000 in Lecture Notes in Computer Science, pages 536–548. Springer-Verlag, 1995.

[58] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.