A Probabilistic Terminological Logic for Modelling Information Retrieval*

Fabrizio Sebastiani†

Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
Via S. Maria, 46 - 56126 Pisa (Italy)
E-mail: fabrizio@iei.pi.cnr.it

Abstract

Some researchers have recently argued that the task of Information Retrieval (IR) may successfully be described by means of mathematical logic; accordingly, the relevance of a given document to a given information need should be assessed by checking the validity of the logical formula $d \rightarrow n$, where d is the representation of the document, n is the representation of the information need and "$\rightarrow$" is the conditional connective of the logic in question. In a recent paper we have proposed *Terminological Logics* (TLs) as suitable logics for modelling IR within the paradigm described above. This proposal, however, while making a step towards adequately modelling IR in a logical way, does not account for the fact that the relevance of a document to an information need can only be assessed up to a limited degree of certainty. In this work, we try to overcome this limitation by introducing a model of IR based on a *Probabilistic* TL, i.e. a logic allowing the expression of real-valued terms representing probability values and possibly involving expressions of a TL. Two different types of probabilistic information, i.e. *statistical information* and *information about degrees of belief*, can be accounted for in this logic. The paper presents a formal syntax and a denotational (possible-worlds) semantics for this logic, and discusses, by means of a number of examples, its adequacy as a formal tool for describing IR.

1 Introduction

In recent years, researchers in Information Retrieval (IR) have devoted an increasing amount of work to the search for *models* of IR, i.e. for theoretical descriptions of the IR process that could serve both as specifications for building running systems, and as theoretical tools for abstractly investigating the relative efficiency of systems built along their guidelines.

The attention of researchers seems lately to have concentrated on the so-called *Logical Model*, first introduced by van Rijsbergen [11]. According to the Logical Model, IR may be seen as the task of retrieving, in response to an information need on the part of the user, all the documents that belong to a given document base and that make the formula $d \rightarrow n$ valid (according to the notion of validity of the chosen logic $\mathcal{L}$), where d and n are the representations of the document and of the information need in the language of $\mathcal{L}$, and "$\rightarrow$" is the "conditional" connective of $\mathcal{L}$.

In fact, the Logical Model does not go as far as specifying *which* logic has to be chosen for modelling IR. As a consequence, the key problem of this research paradigm is the selection of an adequate logic for this task; a number of proposals have thus recently appeared that, with varying degrees of success, attempt to instantiate the Logical Model by means of an appropriate logic.

In a recent paper [9], we have argued that a family of logics suitable (at least, at a first approximation) for modelling relevance of documents to information needs along the guidelines of the Logical Model is that of *Terminological Logics* (TLs); we have gone further to propose one such logic (which we have dubbed MIRTL) that we deemed particularly suited to IR purposes.

However, TLs do not deal with *uncertain information* and *statistical information*. Because of this, our model does not make provisions for the fact that the system is not normally able to assess the relevance of a document to an information need with certainty. Actually, van Rijsbergen stresses that, given that the system cannot reasonably expect to determine relevance "objectively", we should think in terms of the the *probability* that the system attributes to d being relevant to n. In the logical model, IR then

*This work has been carried out in the framework of project FERMI 8134 - "Formalization and Experimentation in the Retrieval of Multimedia Information", funded by the European Community under the ESPRIT Basic Research scheme.

†Current address: Department of Computing Science, University of Glasgow, G12 8QQ Glasgow, United Kingdom. E-mail: fabrizio@dcs.glasgow.ac.uk

becomes the task of computing, for each document d, the real number r such that $P(d \rightarrow n) = r$ is valid, and ranking documents in terms of the associated r.

In this paper, we attempt to solve this problem by extending TLs in a *probabilistic* way. Two different types of probabilistic information will be supported in our framework: probabilistic information as *statistical information*, and probabilistic information as *information on the degree of belief* that the system has in other information. Quite obviously, our framework will also allow the combination of these two types of probabilistic information, i.e. it will allow the representation of the degree of belief that the system has in some piece of statistical information.

One interesting result of this approach is that the resulting logic allows the expression of conditional probabilities involving expressions of a TL; it will then be possible, as advocated by van Rijsbergen, to model the probability of relevance of a document to an information need in terms of a conditional probability. But the interesting facet of this is that this will be possible while staying within the confines of classical logic[1], thus relying on a semantic apparatus that is more intuitive and less controversial than that of the conditional logic advocated by van Rijsbergen.

The paper is structured as follows. In order to make it reasonably self-contained, in Section 2 we give a brief introduction to TLs and to the role they play in the model introduced in [9]. In Section 3 we argue that, for IR purposes, it would be useful to endow TLs with primitives for expressing both statistical information and information about degrees of belief. In Section 4 we go on to specify in full formal detail how both types of probability can actually be embedded in TLs; we do this by specifying a formal syntax and a denotational semantics for a language that allows the expression of real-valued terms representing probability values and possibly involving expressions of a TL. We will call the resulting logics *Probabilistic Terminological Logics* (PTLs); we will also see one example of them, i.e. a probabilistic version of MIRTL, that we will call $\mathcal{P}$-MIRTL. Section 5 concludes.

2 An introduction to Terminological Logics and their use in IR modelling

The basic claim of our previous paper [9] is that Terminological Logics:

1. provide a representation language rich enough to accommodate, in an intuitive "object-oriented" syntax, complex descriptions of documents. This language is rich enough to account for the *multifaceted nature* of documents, i.e. for the fact that documents have a number of "orthogonal" properties (such as content, structure, graphical characteristics, etc.) that users might want to use in referring to them within queries;
2. can accommodate, *in the same representation language, and with the same intuitive "object-oriented" syntax*, descriptions of user information needs complex enough to address the above mentioned multifaceted nature of documents;
3. can accommodate, *in the same representation language, and with the same intuitive "object-oriented" syntax*, the expression of "lexical" information, i.e. the kind of information that IR systems usually store in thesauri;
4. provide for an interesting, semantically clean, and natural way of thinking of the relevance of a document to an information need in terms of the "$d \rightarrow n$ view" put forth by van Rijsbergen.

The primary syntactic expressions of TLs are *terms*. In TLs a term is an expression that denotes either an individual, or a unary or binary relation on the domain of discourse. Terms denoting individuals are called *individual constants* (hereafter indicated by metavariables i, i_1, i_2, ...), while terms denoting unary relations are called *concepts* (indicated by metavariables C, C_1, C_2, ...) and terms denoting binary relations are called *roles* (indicated by metavariables R, R_1, R_2, ...). In the same manner as complex sentences of classical sentential logics are formed by the recursive application of connectives to sentential letters, complex terms of TLs are formed by the recursive application of *term-forming operators* to individual constants, unary predicate symbols (indicated by metavariables M, M_1, M_2, ...) and binary predicate symbols (indicated by metavariables D, D_1, D_2, ...). Each TL has its own set of operators; the one included in MIRTL are detailed in Footnote 2.

In the model of IR developed in [9], a document is represented by an individual constant, while a class of documents is represented by a concept. For example, the individual constant `paper666` might represent a particular document contained in the document base under consideration. Given the unary predicate symbols `paper` and `tl`, the binary predicate symbols `author`, `appears-in`, `affiliation` and `deals-with`, and the individual constants `SIGIR93` and `IEI-CNR`, the expression

```
(and paper
     (func appears-in (sing SIGIR93))
```

[1] TLs may in fact be regarded as fragments of first-order logic.

(all author (func affiliation (sing IEI-CNR)))
(c-some deals-with t1))

is a concept that (under the obvious interpretation for predicate symbols and individual constants, and under the interpretation of term-forming operators detailed in [9]) denotes the class of all those papers that appear in the SIGIR93 proceedings, all of whose authors are affiliated with IEI-CNR, and that deal with TLs[2]. Note that, in describing this class of documents, we have freely interspersed "contextual" information about these documents (i.e. their authors, the affiliation of these authors, the volume in which the papers appear) and information about their "semantic content" (i.e. what these papers deal with); quite obviously, different types of information have the same semantic importance, as they are conveyed in the same, uniform language.

Terminological Logics also allow for *instance assertions* (or simply *assertions* – indicated by metavariables γ, γ_1, γ_2, ...), by means of which one can state that a given individual constant is an instance of a given concept (or that a pair of individual constants is an instance of a given role). In the model of IR developed in [9], an assertion is used to represent the membership of a document with a class of documents, and hence to *describe* the document itself; for instance, the assertion

(and paper
(func appears-in (sing SIGIR93))
(all author (func affiliation (sing IEI-CNR)))
(c-some deals-with t1)) **[paper666]**

states that **paper666** is a paper that appears in the SIGIR93 proceedings, all of whose authors are affiliated with IEI-CNR, and that deals with TLs. The assertion mechanism also allows to represent the fact that a document may itself be a part of a larger document, as in the following example:

(and conference-proceedings
(all (inv appears-in) (c-some deals-with (sing IR)))) **[SIGIR93]**

This assertion states that SIGIR93 is a volume containing the proceedings of a conference, and that all of the papers appearing therein have information retrieval as their common topic. Note, incidentally, that this states something about **paper666** too, because in the previous assertion it had been stated that **paper666** appears in SIGIR93; about **paper666** we now also know that it deals with information retrieval.

TLs also allow stating (by means of *axioms* – indicated by metavariables δ, δ_1, δ_2, ...)) that either a relation of "conceptual containment" ($\sqsubset$) or of "conceptual equivalence" ($\doteq$) holds between two terms. In the model of IR developed in [9], axioms are used to make lexical, "thesaural" knowledge available to the reasoning mechanism in a transparent way. For example, the expression

t1 $\doteq$ **(and logic**
(func syntax term-oriented-syntax)
(func semantics extensional-semantics))

defines TLs to be logics with a term-oriented syntax and an extensional semantics.

It is now easy to see how issues 1÷3 are addressed within the model we have proposed: individual documents are represented by individual constants, and are described by stating, by means of assertions, their membership with classes of documents, which, in turn, are represented by means of concepts. Also information needs, given that they can be viewed as identifying sets of documents (i.e. the set of documents that the user states to correspond to his/her information need), are represented by concepts.

[2] For reasons of space we will not give the formal semantics of the term-forming operators of MIRTL here; the interested reader is referred either to [9] or to the full paper [10]. The *informal* meaning of these operators is the following:

- **(top)** and **(bottom)** denote the set of all individuals of the domain of discourse and the empty set, respectively;
- **(a-not** M**)** denotes the set of all individuals of the domain that are not denoted by M;
- **(sing** i**)** denotes the set containing only the individual denoted by i; this construct (its name standing for "singleton") is included in order to be able to have individual constants as sub-components of concepts;
- **(and** C_1 C_2 ... C_n**)** denotes the set of those individuals that are denoted by C_1 and, at the same time, by C_2 and ... C_n;
- **(all** R C**)** denotes the set of those individuals whose R's are all C's; for instance, **(all author italian)** denotes the set of individuals whose **authors** are all **italians**;
- **(c-some** R C**)** denotes the set of those individuals having at least one R that is a C; for instance, **(c-some author italian)** denotes the set of individuals that have at least one **author** who is an **italian**;
- **(atleast** n R**)** (resp. **(atmost** n R**)**) denotes the set of those individuals having at least (resp. at most) n R's;
- **(inv** R**)** denotes the set containing the inverses of those pairs that are denoted by R; for instance, **(inv husband)** will be, under the obvious interpretation, equal to the role **wife**.

We will also use the following shorthands:

- **(exactly** n R**)** will be used in place of **(and (atleast** n R**) (atmost** n R**))**;
- **(func** R C**)** will be used in place of **(and (all** R C**) (exactly** 1 R**))**.

The only issue we have not yet hinted to is how to accommodate the matching function between queries and documents. In the model described in [9] this was done via the notion of *subsumption* (i.e. "hierarchical domination") between terms. For better compatibility with what is to come in the rest of this paper, we can restate the result in terms of the notion of *validity*, familiar from classical logic: the terminological model sees IR as the task of retrieving, as a response to a query C, all and only those documents i such that $C[i]$ is valid in Ω, i.e. such that the denotation of i belongs to the denotation of C in all interpretations satisfying the *knowledge base* Ω, i.e. the set of assertions describing documents and their membership with document classes and axioms describing thesaural knowledge. In the case of TLs, then, the "instance assertion" operator "[ ]" plays the role of van Rijsbergen's "$\rightarrow$" conditional connective.

3 The need for probabilities

In the previous section we have described how van Rijsbergen's "$d \rightarrow n$" model can be instantiated in the case of TLs. It now remains to be seen how we can extend this model in order to account for van Rijsbergen's recommendation that the logical model of IR should hinge on the computation of the *probability* that $d \rightarrow n$. In order to tackle this problem in our model, we should endow our representation language with the possibility of attributing probabilities to the expressions that are to describe documents, lexical knowledge and queries, and upon which the computation is to be based.

There are at least two different senses in which probabilities are used in IR. One is concerned with the *degrees of belief* (or *degrees of confidence*) that the system (or the indexer) "subjectively" has in some facts, such as the fact that a given document might be relevant to a given information need on the part of the user. The second is concerned with "objective" *statistical* information that the system (or indexer) has, and that is brought to bear in the decision process.

3.1 Information about degrees of belief

The first issue to take into consideration is the possiblity of attributing a probability to an assertion. For instance, the indexer (whether human or machine) might judge that the likelihood that **paper666** would be deemed relevant by a user formulating his/her information need by means of the query

(and paper (c-some deals-with tl))

is greater than or equal to 0.8. In doing this, the indexer is expressing his/her *degree of belief* in the relevance of the document to the query. Accordingly, the representation language of our logic will allow the expression of formulae such as

w(**(and paper (c-some deals-with tl))** [**paper666**]) ≥ 0.8

and, in general, of formulae of type ($w(\gamma)$ *relop* t), where γ is an assertion, t is an expression which evaluates to a real number, and *relop* is one of the relational operators "$=$", "$\neq$", "$\leq$", "$\geq$", "$<$" and "$>$".

Suppose now we want to express the fact that, if we believed that **paper666** is a document that deals with logics, our degree of belief in the fact that **paper666** deals with TLs would be greater or equal than 0.8. This is actually a *conditional probability* notion. But once we have the possibility of representing the degree of belief in an assertion, this is easily expressible, as what this actually means is

$$w(\textbf{and document (c-some deals-with tl) (c-some deals-with logic)}) \geq 0.8 \cdot w(\textbf{and document (c-some deals-with logic)})$$

In keeping with the usual notation for conditional probabilities, we will abbreviate the latter formula as

$w_{(x)}$(**(c-some deals-with tl)** | **(and document (c-some deals-with logic))**) ≥ 0.8

In semantic terms, along with [4], we will model degrees of belief by relying on a "possible worlds semantics" (PWS) [7], i.e. by viewing the different states of affairs (or "worlds", in PWS jargon) that the system considers in principle possible, as grouped into structures. We will postulate the existence of a probability distribution on the set of worlds belonging to a given structure; in a given world, the formula $w(\gamma)$ *relop* t will be true just in case the probability of the set of worlds belonging to the same structure and in which γ is true "relop's" the value of t at that world. Hence, the formula in the above example will be true at world x just in case the probabilities of the worlds which are possible relative to x (i.e. that belong to the same structure as x), and in which **paper666** is indeed a paper that deals with TLs, sum up to at least 0.8.

It goes without saying that the same mechanism can also be used for representing the degree of belief of the system in an axiom. Our representation language will then allow the expression of formulae such as

$w(\mathtt{tl} \sqsubset (\mathbf{and}\ \mathtt{logic}\ (\mathbf{func}\ \mathtt{syntax}\ \mathtt{term\text{-}oriented\text{-}syntax}))) \geq 0.8$

It should be noted that, given the possible-worlds interpretation we have given above, this actually means that the degree of belief that the system has in the fact that *all* TLs have a term-oriented syntax is at least 0.8; it does *not* mean, instead, that the system firmly believes that more than 80% of all TLs have a term-oriented syntax! There is a fundamental difference in these two statements, to the extent that they should be seen as encoding *two different kinds of knowledge*: the truth of the latter statement depends on the objective state of the world, but the truth of the former depends instead on the subjective state of a cognitive agent.

Indeed, the possibility to express a sentence of the latter, "statistical" type is of paramount importance in a model of IR, as statistical notions play a central role in IR. In order to do this, however, we have to introduce further semantic mechanisms.

3.2 Statistical information

As we said in the previous section, the possibility of expressing statistical information would prove of paramount importance in our model. For instance, the indexer might want to express the fact that 80% of the documents in the document base under consideration deal with computer science.

In order to do this, however, we have first to introduce the notion of assigning a probability to a concept (or to a role). By "probability of a concept C" (resp. of a role R) here we mean the probability that a randomly picked individual i (resp. pair of individuals $\langle i_1, i_2 \rangle$) turns out to be a C (resp. a R). This interpretation is in keeping with the fact that here we want to model a *statistical* notion of probability: we intend to describe either a chance setup, or the statistical information that we might possess as a result of experimentation on this chance setup. The real-valued quantity "the probability that a randomly picked individual is a C" (resp. that a pair of randomly picked individuals are a R) will be expressed by the probability term $w_{\langle x \rangle} C$ (resp. by the probability term $w_{\langle x_1, x_2 \rangle} R$).

Suppose now we want to express the fact that 80% of the documents in the document base under consideration deal with computer science. What this actually means is that the probability that an individual which has been randomly picked out of the set of documents is about computer science is 0.8. This is, again, a conditional probability notion, which is easily expressible, as it may be written as

$w_{\langle x \rangle}(\mathbf{and}\ \mathtt{document}\ (\mathbf{c\text{-}some}\ \mathtt{deals\text{-}with}\ (\mathbf{sing}\ \mathtt{CS}))) = 0.8 \cdot w_{\langle x \rangle}(\mathtt{document})$

Similarly to the case of degrees of belief, we will abbreviate the latter formula as

$w_{\langle x \rangle}((\mathbf{c\text{-}some}\ \mathtt{deals\text{-}with}\ (\mathbf{sing}\ \mathtt{CS})) \mid \mathtt{document}) = 0.8$

In semantic terms, we will have to take a radically different approach from the one based on "possible worlds" of the previous section. Here we do not want to model the system's uncertainty about the truth value of a (non-probabilistic) proposition; rather, we want to model the system's certainty about the truth value of a probabilistic proposition! What we need here is to "stay in a fixed world", and postulate a probability distribution on the domain of discourse of that world: the probability of a concept will then be the sum of the probabilities of the elements in the domain that belong to the denotation of that concept. If the probability distribution is *uniform*, this corresponds to checking the *relative cardinality* of the concept, i.e. the percentage of individuals of the domain that belong to the denotation of that concept [1]. For instance, the probability of the concept (**c-some** `deals-with` (**sing** `CS`)) will be the sum of the probabilities of those individuals that belong to its denotation, i.e. that deal with computer science.

4 A Probabilistic Terminological Logic

In the previous section we have informally argued how the introduction of probabilities into our "terminological model" would allow the expression of a number of notions interesting to IR. In this section we proceed to specify in full detail the syntax and semantics of a particular PTL, that we have dubbed $\mathcal{P}$-MIRTL. Actually, $\mathcal{P}$-MIRTL consists of an extension of the MIRTL logic by means of probabilistic features. It goes without saying that $\mathcal{P}$-MIRTL should serve only as an example of the potentialities of PTLs in modelling IR; other TLs different from MIRTL might be "plugged" in the framework without any added work.

The semantics of our logic will be given in the style of *denotational semantics*. For the logically uninitiated, we should say that denotational semantics (also known as *model-theoretic* or *Tarskian semantics*) is the standard way of formally specifying the meaning of logical languages. Such a specification is accomplished by postulating the existence of a number of "ways the world could be" (*interpretations*), and of systematically specifying in which of these interpretations the expressions of the language are true; "systematically" here means that the semantic specification mirrors the recursive structure of the syntactic BNF specification, with one semantic clause for each syntactic clause. Inference is then defined

as the derivation of only those formulae that are true in all the interpretations in which the premises are also true. The specification that follows fully conforms to this systematic pattern.

For reasons of space, we will not specify the semantics of MIRTL in full detail, but will limit our discussion to its probabilistic extension; the interested reader may consult either [9] or the full paper [10].

4.1 Syntax

The syntax of $\mathcal{P}$-MIRTL extension hinges on the notions of "formula" and "probability term". These will be defined in a mutually recursive way.

Definition 1 *A* probability term *is either a rational constant z, or is an expression of the form $w_{\langle x\rangle}C$, or of the form $w_{\langle x_1,x_2\rangle}R$, or of the form $w(\phi)$ (with ϕ a* formula*), or of the form $(t_1$ mathop $t_2)$, where both t_1 and t_2 are probability terms and "mathop" is an operator in the set $MATHOP=\{+,-,\cdot,\div\}$.*

A formula *is either an axiom δ, or an assertion γ, or an expression of the form $(t_1$ relop $t_2)$, where both t_1 and t_2 are probability terms and "relop" is an operator in the set $RELOP=\{=,\neq,\geq,\leq,<,>\}$.*

We will use metavariables $\phi, \phi_1, \phi_2, \ldots$ ranging on formulae and metavariables $t, t_1, t_2, \ldots$ ranging on probability terms. As hinted in Section 3, we will use the notation $w_{\langle x\rangle}(C_1|C_2)$ *relop* t as shorthand for the expression $w_{\langle x\rangle}$(**and** C_1 C_2) *relop* $(t \cdot w_{\langle x\rangle}C_2)$, and the notation $w(C_1[i]|C_2[i])$ *relop* t as shorthand for the expression w(**and** $C_1[i]$ $C_2[i]$) *relop* $(t \cdot w(C_2[i]))$.

Note that our language is indeed much more powerful than the examples of Section 3 show[3]. For instance, in formulae of type $(t_1$ *relop* $t_2)$, term t_2 is not restricted to be a numeric expression; it is therefore possible to compare, by means of a relational operator, two "complex" probability terms, as in

$$w_{\langle x\rangle}(\texttt{and document (c-some deals-with (sing CS))}) < w_{\langle x\rangle}(\texttt{and document (c-some deals-with (sing Mathematics))})$$

Also, it is possible to "nest" probability operators, e.g. to express how strongly the system believes in some proposition of a statistical nature. For example, it is possible to write two formulae like

$$w(w_{\langle x\rangle}(\texttt{multimedia-document} \mid \texttt{document}) \geq 0.7) = 0.1$$

$$w(w_{\langle x\rangle}(\texttt{multimedia-document} \mid \texttt{document}) \geq 0.2) = 0.8$$

whose combined effect is to assert that the system is not inclined to believe (i.e. its degree of belief is 0.1) that more than 70% of the documents in the collection are multimedia documents, but is definitely more inclined to accept (i.e. its degree of belief is 0.8) that this percentage might only be above 20%.

4.2 Semantics

Now that we have completely detailed the syntax of the probabilistic features of $\mathcal{P}$-MIRTL, we may switch to discussing its semantics (a semantics that will follow the guidelines of Halpern's $\mathcal{L}_3$ logic [4]). As previously hinted, a denotational semantics for a logical language is obtained by postulating the existence of a number of "ways the world could be"; these are usually called *interpretations*. In our case, these will exactly be the "interpretations" of MIRTL as defined and characterised in Definitions 1÷5 of [9]. Such interpretations consist of mappings of individual constants into individuals of the domain, and of predicate symbols into relations on the domain, that are "well-behaved" with respect to the intuitive meaning of the operators of the language (i.e. the term-forming operators, the assertion operator "[ ]" and the axiom operators "$<$" and "$\doteq$".).

In order to give semantics to the probabilistic features of $\mathcal{P}$-MIRTL, we will adopt a version of "possible world semantics" (PWS); as in all versions of PWS, we will see the set of interpretations as partitioned into structures, that we will call *PTL structures*[4]. A PTL structure is a 4-uple $M = \{\mathcal{D}, I, \nu_{dom}, \nu_{int}\}$, where $\mathcal{D}$ is a nonempty set of individuals, I is a set of MIRTL interpretations on $\mathcal{D}$, ν_{dom} is a discrete probability distribution on the domain $\mathcal{D}$, and ν_{int} is a discrete probability distribution on the set I of interpretations.

[3] In the full paper [10] we also consider formulae of type $w_{\langle x_1,x_2\rangle}R$ where either x_1 or x_2 may be an individual constant. This allows the expression of probability terms such as $w_{\langle x,\text{SIGIR93}\rangle}$(**appears-in**), which expresses the "probability that a randomly picked x appears in the SIGIR93 proceedings".

[4] In most approaches based on PWS, interpretations are called "possible worlds"; we avoid using this terminology here, both because it has often given rise to misunderstandings about the meaning of "possible", and because we want to highlight the relationship of containment between the semantics of MIRTL and the semantics of $\mathcal{P}$-MIRTL. It is worthwhile to notice that, in the same way that possible worlds for e.g. sentential modal logic are truth value assignments (to the formulae of sentential logic) that comply with the intuitive meaning of the connectives of sentential logic, interpretations are assignments of "extensions" (to concepts, roles and individual constants) that comply with the intuitive meaning of the operators of MIRTL. It is thus clear that "classical" (sentential) possible worlds stand to sentential logic as our interpretations stand to MIRTL.

The notion of *extension* of a probabilistic term t in an interpretation $\mathcal{I}$ of M, written $[t]_{(M,\mathcal{I})}$, and the notion of *truth* of a formula ϕ in an interpretation $\mathcal{I}$ of M, written $(M,\mathcal{I}) \models \phi$, are defined, in a mutually recursive way (this is obvious, given that the syntax of probabilistic terms and formulae is also mutually recursively defined), by means of the following clauses:

1. $[z]_{(M,\mathcal{I})} = \mathbf{z}$
2. $[w_{\langle x \rangle} C]_{(M,\mathcal{I})} = \sum_{d \in \mathcal{I}(C)} \nu_{dom}(d)$
3. $[w_{\langle x_1, x_2 \rangle} R]_{(M,\mathcal{I})} = \sum_{(d_1,d_2) \in \mathcal{I}(R)} \nu_{dom}(d_1) \cdot \nu_{dom}(d_2)$
4. $[w(\phi)]_{(M,\mathcal{I})} = \sum_{\mathcal{J} \in I \,:\, (M,\mathcal{J}) \models \phi} \nu_{int}(\mathcal{J})$
5. $[t_1 \; mathop \; t_2]_{(M,\mathcal{I})} = [t_1]_{(M,\mathcal{I})}$ **mathop** $[t_2]_{(M,\mathcal{I})}$
6. $(M,\mathcal{I}) \models \delta$ iff $\mathcal{I}$ satisfies δ
7. $(M,\mathcal{I}) \models \gamma$ iff $\mathcal{I}$ satisfies γ
8. $(M,\mathcal{I}) \models t_1 \; relop \; t_2$ iff $[t_1]_{(M,\mathcal{I})}$ **relop** $[t_2]_{(M,\mathcal{I})}$

We will now comment on the meaning of each of these clauses and then give a comprehensive example of their use. Clause 1 simply states that the extension of a rational constant is always (i.e. in any interpretation $\mathcal{I}$ of any PTL structure M) the rational number it obviously represents (e.g. the constants 0.25, $\frac{1}{4}$ and $\frac{4}{16}$ all represent the number **0.25**). Analogously, Clause 5 states that the extension of a term $(t_1 \; mathop \; t_2)$ is always obtained by the application of the real-valued operation **mathop** which the symbol "*mathop*" obviously represents (e.g. the "+" symbol representing addition), to the extensions of the two terms involved; similar considerations apply to Clause 8.

Probabilities come in with Clause 2: in order to compute the extension, at a given interpretation $\mathcal{I}$ of M, of the probability that a randomly picked x be a C, we first check what individuals belong to the interpretation of C under $\mathcal{I}$, and then sum up the probabilities that the distribution ν_{dom} attributes to them. The case of Clause 3 is completely analogous, the only difference being that pairs of individuals (and, consequently, the product of their probabilities) have to be considered instead of single individuals.

Things are quite different with Clause 4; this clause aims to specify the semantics of formulae involving the system's degrees of belief, so a reference must be made to the interpretations that the system "believes in principle possible" and to their respective probabilities. In order to compute the system's degree of belief in a formula, we first check what are the interpretations in which that formula is true, and then sum up the probabilities that the distribution ν_{int} attributes to them[5].

Finally, things are quite simple for Clauses 6 and 7; MIRTL assertions and axioms are true in an interpretation $\mathcal{I}$ just if they are satisfied by $\mathcal{I}$ in the sense of Definitions 2 and 3 of [9].

Let us now work out a simple example in order to see how all this works.

Example 1 *Suppose we have a PTL structure* $\mathcal{M} = \{\mathcal{D}, I, \nu_{dom}, \nu_{int}\}$, *where:*

- $\mathcal{D}$ *is the set containing the three individuals* a, b *and* c;
- I *is the set consisting of the two interpretations* $\mathcal{I}_1$ *and* $\mathcal{I}_2$, *which, given the two unary predicate symbols* **d** *and* **md** *(which we might take to stand for "document" and "multimedia document", respectively), are such that* $\mathcal{I}_1(\mathtt{d}) = \{a, b\}$, $\mathcal{I}_1(\mathtt{md}) = \{a, b\}$, $\mathcal{I}_2(\mathtt{d}) = \{a, b, c\}$ *and* $\mathcal{I}_2(\mathtt{md}) = \{a\}$;
- ν_{dom} *is a discrete probability distribution on* $\mathcal{D}$ *such that* $\nu_{dom}(a) = 0.1$, $\nu_{dom}(b) = 0.3$ *and* $\nu_{dom}(c) = 0.6$;
- ν_{int} *is a discrete probability distribution on* I *such that* $\nu_{int}(\mathcal{I}_1) = 0.65$ *and* $\nu_{int}(\mathcal{I}_2) = 0.35$.

Suppose we want to know what is the extension of the term

$$w(w_{\langle x \rangle}(\mathtt{md}|\mathtt{d}) \geq 0.8)$$

(i.e. what is the system's degree of belief in the fact that 80% of the documents in the collection are multimedia) in interpretation $\mathcal{I}_1$. *This term is shorthand for the term*

[5] Readers familiar with PWS for modal logic might have noticed that, unlike what normally happens for formulae whose interpretation depends on a multiplicity of "possible worlds", there is no "accessibility relation" involved in the computation of degrees of belief. This is not inconsistent with the principles underlying PWS, as it is well-known (see e.g. [5, pages 334–335]) that using the set of all worlds belonging to a modal structure is equivalent to using only the set of worlds that are "accessible" through an equivalence relation. $\mathcal{P}$-MIRTL is then conceptually similar to the S5 modal logic; analogously to what happens in S5, terms representing degrees of belief have the same extension in all interpretations belonging to the same PTL structure.

$w(w_{\langle x\rangle}(\textbf{and md d}) \geq 0.8 \cdot w_{\langle x\rangle}(\textbf{d}))$

According to Definition 1 of [9], $\mathcal{I}_1((\textbf{and md d})) = \{a, b\}$ *and* $\mathcal{I}_2((\textbf{and md d})) = \{a\}$. *It follows that* $w_{\langle x\rangle}(\textbf{and md d})$ *evaluates to* $\nu_{dom}(a) + \nu_{dom}(b) = 0.4$ *in* $\mathcal{I}_1$, *and to* $\nu_{dom}(a) = 0.1$ *in* $\mathcal{I}_2$. *Given that* $w_{\langle x\rangle}(\textbf{d})$ *obviously evaluates to* $\nu_{dom}(a) + \nu_{dom}(b) = 0.4$ *in* $\mathcal{I}_1$, *and to* $\nu_{dom}(a) + \nu_{dom}(b) + \nu_{dom}(c) = 1$ *in* $\mathcal{I}_2$, *the formula*

$w_{\langle x\rangle}(\textbf{and md d}) \geq 0.8 \cdot w_{\langle x\rangle}(\textbf{d})$

is true in $\mathcal{I}_1$ *but false in* $\mathcal{I}_2$. *Hence, the extension of the term*

$w(w_{\langle x\rangle}(\textbf{md}|\textbf{d}) \geq 0.8)$

in $\mathcal{I}_1$ *is equal to* $\nu_{int}(\mathcal{I}_1) = 0.65$.

Notice that this term has the same extension also in interpretation $\mathcal{I}_2$; *as we have already noticed in Footnote 5, terms expressing degrees of belief evaluate to the same value in all interpretations belonging to the same PTL structure.*

Similarly to what happens in all applications of logical reasoning, we are hardly interested in what is the truth value of a given formula, or the extension of a given term, at a *particular interpretation* $\mathcal{I}$; loosely speaking, we cannot know which interpretation is *the* correct one, i.e. the one that corresponds to the "real world", since we always have *partial* (and often *erroneous*) knowledge about the real world (in our case: about the documents in our collection and about what they are about), and much of what is true in the real world is unknown to us. Because of this, we are rather interested in what is the truth value of a given formula, or the extension of a given term, at all those interpretations that are "consistent" with our partial and erroneous knowledge about the world; this corresponds to the logical process of *inferring* those formulae whose truth is a consequence of the truth of the formulae that constitute our knowledge about the world. As in all other logics, in $\mathcal{P}$-MIRTL this is formalised by the notion of *validity in a theory.*

Definition 2 *A* theory *of* $\mathcal{P}$-MIRTL *is a set* Φ *of formulae which is closed under logical consequence; i.e.* Φ *is such that, if* $\phi \in \Phi$ *and* ϕ' *is true in all the interpretations in which* ϕ *is true, then also* $\phi' \in \Phi$.

Definition 3 *A formula* ϕ *is* valid *in a theory* Φ *of* $\mathcal{P}$-MIRTL, *written* $\models_\Phi \phi$, *iff* ϕ *is true in all interpretations* $\mathcal{I}$ *(of all PTL structures* M*) in which all formulae in* Φ *are also true.*

Note that the above observations on partial and erroneous knowledge apply to probability distributions too. By relying on the notion of validity in a theory Φ, we free ourselves from the problem of knowing, in all details, which probability distribution on the domain (resp. on possible worlds) is *the* correct one. This is reasonable, as we could not hope to know the truth value of every (probabilistic) formula expressible in our language: also our *probabilistic* knowledge is partial, and often erroneous too! A set of formulae Φ does not specify a probability distribution in full detail, but has the effect of putting a number of constraints on how probability distributions "consistent" with Φ should be; these constraints identify a whole family of distributions, and the formulae valid in Φ are exactly those formulae that are true in all the interpretations characterised by these distributions.

An interesting side-effect of introducing the notion of probability distribution into the semantics of our logic is that our logic will obey the familiar laws of the probability calculus; this will be true both for formulae representing information about degrees of belief, and for formulae representing statistical information. For example, Bayes' Theorem is valid in our logic, i.e. all formulae of type

$$w_{\langle x\rangle}(C_1|C_2) \cdot w_{\langle x\rangle}(C_2) = w_{\langle x\rangle}(C_2|C_1) \cdot w_{\langle x\rangle}(C_1)$$

or of type

$$w(C_1[i]|C_2[i]) \cdot w(C_2[i]) = w(C_2[i]|C_1[i]) \cdot w(C_1[i])$$

will be valid in any theory Φ, as can easily be seen by applying our definition of conditional probability.

According to the model of IR that we are proposing in this paper, a document i is then deemed to be relevant to an information need C with probability r, with r a real number, iff the formula $w(C[i]) = r$ is valid in Φ, where Φ is the (consequential closure of) the set of formulae representing the documents in the collection and the lexical, "thesaural" knowledge of the system.

5 Concluding remarks

In this paper we have presented a logical model for information retrieval based on a probabilistic terminological logic. In this model, IR is seen as the task of 1) computing, for a given information need (represented by the concept) C and for each document (represented by an individual constant) i, the real number (represented by the constant) r such that $w(C[i]) = r$ is valid in Φ (i.e. in the theory representing the document base and the lexical, "thesaural" knowledge), and 2) ranking documents in terms of their associated r.

Besides enjoying the numerous properties that accrue from the adoption of a TL (properties that are more fully described in [9]), this model takes advantage of the considerable expressive power provided by our probabilistic extension to the terminological framework. This extension allows the distinct expression of two radically and conceptually different kinds of probabilistic information that feature in the IR task, i.e. *statistical* information, and information about the *degrees of belief* that the IR system being modelled has in other information.

Although statistical information and information about degrees of belief are conceptually different, it is clear that there is a relationship between the two. Our work so far has aimed at providing a framework in which both could be expressed and reasoned upon in a principled, semantically clear way. A further step in this direction should be the investigation of mechanisms for allowing information about degrees of belief to be directly *derivable* from statistical information. For instance, ifthe system has no belief at all (i.e. to no degree) whether a given assertion $C[i]$ is true, but at the same time knows that 80% of all individuals of the domain are C's, it might plausibly decide to believe with a 0.8 degree of confidence that i, a particular individual in the domain, is a C. This approach to the derivation of degrees of belief, well known in actuarial reasoning, is known as *direct inference* (see e.g. [8]). Other approaches exist however, yielding different results, and based on principles as diverse as the *maximum entropy principle* (see e.g. [3]), the *centre of mass principle* or the *maximal independence principle* (see e.g. [2]). Unfortunately, in all of these approaches, degrees of belief are *completely determined* by statistical information, to the extent that two formulae such as $w(C[i]) = r_1$ and $w_{\langle x \rangle}(C) = r_2$ would jointly imply that $r_1 = r_2$; instead, it is clear that we would like to be able to entertain such beliefs without this implying that $r_1 = r_2$. Investigating mechanisms that allow statistical information to determine degrees of belief *only when these latter are not already determined* is the next research task that this work opens up.

Acknowledgements

I would like to thank Mark Sanderson for reading an earlier draft and providing useful comments.

References

1. F. Bacchus. *Representing and reasoning with probabilistic knowledge.* MIT Press, Cambridge, MA, 1990.
2. F. Bacchus, A. Grove, J. Y. Halpern, and D. Koller. From statistics to beliefs. In *Proceedings of AAAI-92, 10th Conference of the American Association for Artificial Intelligence*, pages 602–608, San Jose, CA, 1992.
3. A. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. In *Proceedings of the 7th annual IEEE Symposium in Logic in Computer Science*, pages 22–33, Santa Cruz, CA, 1992.
4. J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.
5. J. Y. Halpern and Y. O. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.
6. W. L. Harper, R. Stalnaker, and G. Pearce, editors. *Ifs. Conditionals, belief, decision, chance and time.* Reidel, Dordrecht, NL, 1981.
7. S. A. Kripke. Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67–96, 1963.
8. H. Kyburg. The reference class. *Philosophy of science*, 50:374–397, 1983.
9. C. Meghini, F. Sebastiani, U. Straccia, and C. 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.
10. F. Sebastiani. A model of information retrieval based on a probabilistic terminological logic (extended version). Technical report, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy, 1994. Forthcoming.
11. C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.
12. C. J. van Rijsbergen. Probabilistic retrieval revisited. *The Computer Journal*, 35:291–298, 1992.