

Fusing syntax and semantics in knowledge fusion

É. Grégoire

CRIL Université d'Artois
Rue de l'Université
F-62307 Lens Cedex France
gregoire@cril.univ-artois.fr

A. Sofiane

CRIL Université d'Artois
Rue de l'Université
F-62307 Lens Cedex France
sofiane@cril.univ-artois.fr

Abstract

In this paper, a series of knowledge fusion operators are motivated and analyzed. They are defined in a semantic way, although syntactical facets of knowledge are taken into account. More precisely, they rely on a rank-ordering of interpretations that is based on the number of formulas that the interpretations falsify. It is briefly discussed how these operators could be refined, by taking into account various distribution policies of the falsified information among the knowledge sources, syntactical properties of formulas to be fused and forms of integrity constraints preference among literals.

Keywords: knowledge fusion, knowledge representation

1 Introduction

In this paper, we are concerned with fusion operators of logical knowledge or belief bases and discuss a possible trade-off that can be adopted between purely semantical and syntactical points of view. Indeed, when logic is used to represent knowledge, two dual facets of logic must be dealt with: syntax and model-theoretic semantics, which correspond to two different levels of representation granularity. As Levesque emphasized it [10], standard model-theoretic semantics is too coarse-grained whereas a purely syntax-based view of logic can be too fined-grained. For example, from a model-theoretic point of view, there is no difference between logically equivalent formulas, whereas at the other extreme both formulas $a \wedge b$ and $b \wedge a$ can be seen as different ones from a purely syntactical point of

view. Accordingly, in many logic-oriented knowledge representation issues, tuning the expressiveness granularity according to the application needs is a key problem.

In this paper, such an issue is addressed in the context of the knowledge or beliefs fusion problem, where both model-theoretic-oriented and syntax approaches exist. Fusion operators that offer an interesting trade-off between purely syntactical and semantical points of view are motivated and analyzed. They take syntactical facets of knowledge into account although they are semantically defined. More precisely, they rely on a rank-ordering of interpretations that is based on the number of formulas that the interpretations falsify. It is then shown how these operators could be refined in several ways by taking into account various distribution policies of the falsified information among the knowledge sources, syntactical properties of formulas to be fused and forms of integrity constraints preference among literals. Due to lack of space, the focus is on the motivations and informal description, mainly.

2 Syntactical vs. model-theoretical approaches

For the simplicity of the presentation, we use standard propositional logic, although all results in the paper can be easily extended to the first-order finite Herbrand case. We follow the notations from [1].

Let L be a propositional language of formulas over a finite alphabet P of Boolean variables, also called *atoms*. The \wedge , \vee , \neg and \Rightarrow symbols represent the standard conjunctive, disjunctive, negation and material implication connectives, respectively. A *literal* is an atom or a negated atom. Ω denotes the set of all *interpretations* of L , which are functions assigning either *true* or *false* to every atom. A *model*

of a knowledge base KB is an interpretation that satisfies every formula of KB . An interpretation or a model will be represented by the set of literals that it satisfies. The set of models of KB will be denoted $[[KB]]$. KB is *consistent* when $[[KB]]$ is not empty. $KB \models x$ expresses that the literal x can be deduced from KB , i.e. that it belongs to all models of KB .

Let us consider a multi-set of n ($n > 1$) consistent propositional knowledge bases $E = \{KB_1, \dots, KB_n\}$ to be fused. Fusion operators will be defined as functions Δ that associate a knowledge base, denoted $\Delta(E)$, to each information set E .

From a syntactical point of view, a knowledge base KB is thus a set of formulas of L . Many syntactical approaches to knowledge fusion amount to taking preferred maximal (with respect to cardinality) consistent subset(s) of formulas of the set-theoretic union of the bases to be fused (see e.g. [7, 8]).

From a purely model-theoretical point of view, a KB is a set of models. Most model-theoretic approaches rank order the set of interpretations w.r.t. all KB_i , using the Hamming distance (also called Dalal's distance [3]) from a model of KB_i . $\Delta(E)$ is then characterized by the set of interpretations that are minimal in some sense with respect to this rank-ordering.

In the following, we shall adopt a trade-off between these two points of view. But let us first describe the model-theoretic approach in more details.

The Hamming distance between an interpretation ω and a propositional knowledge base KB_i is defined as the smallest number of atoms about which this interpretation differs from some model of KB_i .

Definition 1.

$d(\omega, KB_i) = \min_{\omega' \in [[KB_i]]} \text{dist}(\omega, \omega')$ where $\text{dist}(\omega, \omega')$ is the number of atoms whose evaluation differs in the two interpretations.

In the following we shall investigate several definitions for rank-ordering interpretations from Ω . Accordingly, assume for the moment that an overall distance noted $d_\Delta(\omega, E)$ between an interpretation and the multi-set E has been defined, already.

Definition 2. $\omega \leq_{(\Delta, E)} \omega'$ iff $d_\Delta(\omega, E) \leq d_\Delta(\omega', E)$

The fused knowledge-base $\Delta(E)$ is defined by its models, which are minimal with respect to $\leq_{(\Delta, E)}$.

Definition 3. $[[\Delta(E)]] = \min(\Omega, \leq_{(\Delta, E)})$

Let us now present the various usual definitions for $d_\Delta(\omega, E)$ that rank-order interpretations.

Definition 4. Majority operators [12, 14]

$$d_\Sigma(\omega, E) = \sum_{(i \in [1..n])} d(\omega, KB_i)$$

Definition 5. Weighted sum operator [11]

$$d_{ws}(\omega, E) = \sum_{(i \in [1..n])} d(\omega, KB_i) * k_i \text{ where } k_i \text{ are numbers reflecting the level of importance of } KB_i.$$

Definition 6. Max-based egalitarian operator [15]

$$d_{max}(\omega, E) = \text{Max}_{(i \in [1..n])} d(\omega, KB_i)$$

In a similar way, lexicographic-based egalitarian operators have been defined in [5-8].

3 Discussion

Let us analyze two main problems with the above semantic fusion operators.

These operators do not take the syntactical form of the knowledge bases to be fused into account; they thus suffer from all the logical omniscience drawbacks [10]. For example, they will deliver the same results if we replace a knowledge base by a logically equivalent one. Whereas some authors claim that this is a necessary requirement, we believe that the syntactical form of the knowledge base *is* to be taken into account, at least in some application domains and to some extent. Indeed, the basic pieces of information introduced by the user in the knowledge base are *formulas*: this should be taken into account in some way. Let us stress that we shall not adopt in this paper an extreme syntactical point of view, which would amount to considering e.g. formulas $a \wedge b$ and $b \wedge a$ as different ones. We shall however justify operators that take the length of formulas into account and can enforce some preferences among literals in the fusion process.

Another problem is that they are not suited to handle inconsistent knowledge bases to be fused (since these latter ones do not exhibit models). This is really a paradox: since they are devised to overcome the situation where the set-theoretic union of the formulas of the knowledge bases to be fused is inconsistent; it would be natural to expect them to be able to handle a basic knowledge base that is itself inconsistent.

In order to introduce formulas as basic ingredients while keeping the above semantical treatment of the fusion process, and in order to solve the above

restriction about handling inconsistent knowledge bases, we just redefine Definition 1 as follows, while we keep Definitions 2 to 6 as such.

Definition 1’.

$d(\omega, KB_i)$ = number of formulas from KB_i that are falsified in the interpretation ω .

First, let us note that this definition applies to inconsistent knowledge bases KB_i as well. Clearly, this definition will strengthen the importance of the syntax of the knowledge bases in the fusion process and will lead to intuitively natural operators.

Clearly, the resulting new *majority* operator coincides with Konieczny’s symmetrical difference operator [7, 8] and delivers a set $[[\Delta(E)]]$ that is the set of models of all the maximal (w.r.t. cardinality) consistent subbases of the set-theoretic union of the bases in E . The resulting new *weighted sum*, *max-based* and *lexicographic-based egalitarian* operators allow us to balance the authorized number of falsified formulas in the different bases. To the best of our knowledge, these latter operators have not yet been described.

The resulting operators are syntax-sensitive also in the sense that they are redundancy-sensitive and in the sense that splitting or merging formulas will affect their results. We claim that this is sometimes a required property. For example, duplicating formulas will directly affect the new operators. This can be justified with respect to some applications, where e.g. a knowledge base is already an accumulation of information from several sources. In this context, duplication of information is sometimes a natural way of enforcing the assertion of the *true* nature of this latter information. Accordingly, dropping this information in a fusion process can be less acceptable than dropping another piece of information that is single-time asserted. Similarly, splitting a conjunctive formulas into its conjuncts will also affect the new distance and operators. This can be justified in several application domains, also. We claim that when a user expresses the formula $a \wedge b \wedge c$, he (she) might want to stress that the three facts a , b and c are crossly-related *true*, which is not necessarily the case when he (she) introduces the three different facts independently in the knowledge base.

4 Distribution of the discarded information

One possible drawback of the new majority, max-based and lexicographic-based operators is that they

do not take the respective size of the different bases to be fused into account. Imagine that we have to fuse a large base with a small one. One way to correct the effect of the size difference between the two bases consists in applying corrective coefficients as in Definition 5, or else, modify Definition 1’ as follows.

Definition 1’.

$d(\omega, KB_i)$ = proportion of formulas from KB_i that are falsified in the interpretation ω .

Let us stress that the corrective factors of Definition 5 and Definition 1’ apply to the base as a whole, and additional corrective factors should be introduced when one wants to take some finer-grained syntactical facets into account.

For example, one could want to take the length of the formulas into account in the distance definition.

In some applications, we accept to falsify basic facts whereas we are less likely to accept it for longer formulas. For example, in the definition of the first expert systems shells *à la* OPS-5, some wired heuristics in the reasoning control system were implementing such a preference. The motivation was twofold. On the one hand, a long formula can represent a more permanent piece of knowledge representing e.g. an uncontroversial rule whereas a single literal is a tentative conclusion or observation that can be contradicted. Secondly, the longer a formula is, the more specific it can be; a more specific piece of information could be preferred because it is more focused and less subject to possible exceptions.

In other domains, the opposite view can be adopted. Facts are uncontroversially *true observations* while rules can be hypothetical and be defeated.

There are many possible ways to take this syntactical facet into account. The more general one would amount to attach numerical factors to formulas in an explicit way and modify the distance concept to take it into account. Obviously enough, such factors could also numerically translate other criteria about the confidence of the piece of information they would be attached to. Accordingly, we could be led to considering general numerical framework such a possibilistic logic [1]. However, when the length of formulas (e.g. the number of literals it is made of) is only to be taken into account, it is possible to adapt the notion of distance directly and avoid the needs of expressing and

representing such explicit numerical factors. Various possible definitions translate various policies in this respect. When counting the number (or proportion) of falsified formulas, we could actually compute for each falsified formula a corresponding function on the length of the formula. Any function could be envisaged, from a constant one which corresponds to not taking the length into account to a function that exponentially increases or decreases with respect to the length of the formula.

Priorities between literals themselves could be taken into account as well. The resulting distances would lead to distances that can be mixed with criteria about the length of formulas. For instance, an extreme policy would require that a given atom, say a , should not appear in any falsified clause. Although, this approach would be very close to enforcing integrity constraints, it is more flexible in the sense that various levels of enforcement can be integrated inside the global distance concept encompassing a weight of the length of the formula itself. All such improvements are discussed in the full version of the paper, together with computational issues.

5 Related works

Some previous works attempted to combine syntactical and semantical facets of knowledge in fusion operators. Several authors, like Konieczny [5], introduced some operators that allow a selection of some preferred maximal consistent sub-bases, using syntactical criteria. Clearly, the operators that we propose do not necessarily lead to maximal consistent sub-bases. In a more general framework, C. Lafarge and J. Lang defined some aggregation functions of preference for group decision making, based on the sum and the max operators [9]. However, they did not consider the lexicographic-oriented operator explicitly.

Acknowledgements

This work has been supported in part by a contract with the Région Nord/Pas-de-Calais. Thanks to S. Konieczny for discussions about the contents of the paper.

References

[1] Benferhat, S., D. Dubois, S. Kaci, H. Prade (2000). Encoding classical fusion in possibilistic logic: a general framework for rational syntactic merging. In: *Proc. of ECAI'2000*.

[2] L. Cholvy (1998). Reasoning about merging information. *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, 3, pp. 233-263.

[3] M. Dalal (1988). Investigations into a theory of knowledge-base revision: preliminary report. In *Proc. AAAI'88*, pp. 475-479.

[4] P. Hansen, Ph. Jaumard (1990). Algorithms for the maximum satisfiability problem. *Journ. of Computing*, 22, pp. 279-303.

[5] S. Konieczny (1999). *Sur la logique du changement: révision et fusion des bases de connaissances*. Thèse d'Université, Univ. de Lille 1.

[6] S. Konieczny, R. Pino Perez (1998). On the logic of merging. In *Proc. of KR'98*, pp. 488-498.

[7] S. Konieczny, R. Pino Perez (1999). Merging with integrity constraints. In *Proc. of Ecsqaru'99*, pp. 233-244, LNCS 1638.

[8] S. Konieczny (2000). On the difference between merging knowledge bases and combining them. In *Proc. of KR'2000*, pp. 135-144.

[9] C. Lafarge, J. Lang (2000). Logical representation of preferences for group decision making. In *Proc. of KR'2000*, pp. 457-468.

[10] H.J. Levesque (1984). A logic of implicit and explicit belief. In *Proc. of AAAI-84*, pp. 198-202.

[11] J. Lin (1996). Integration of weighted knowledge bases, *Art. Int.*, 83, pp. 363-378.

[12] J. Lin, A.O. Mendelson (1998). Merging databases under constraints. *Int. Journ. of Cooperative Information Systems*, 7(1), pp. 55-76.

[13] N. Rescher, R. Manor (1970). On inference from inconsistent premises, *Theory and Decision*, 1, pp. 179-219.

[14] P.Z. Revesz (1993). On the semantics of theory change: arbitration between old and new information. In *Proc. of the 12th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Databases*, pp. 71-92.

[15] P.Z. Revesz (1997). On the semantics of arbitration, *Int. Journ. of Algebra and Computation*, 7(2), pp. 133-160.