

A Probabilistic-Logical Framework for Ontology Matching



Abstract

Ontology matching is the problem of determining correspondences between concepts, properties, and individuals of different heterogeneous ontologies. We present a novel probabilistic-logical framework for ontology matching based on Markov logic. We define the syntax and semantics and provide a formalization of the ontology matching problem within the framework. The approach has several advantages over existing methods such as ease of experimentation, incoherence mitigation during the alignment process, and the incorporation of a-priori confidence values. We show empirically that the approach is efficient and more accurate than existing matchers on an established ontology alignment benchmark dataset.

Markov Logic

Markov logic *combines* first-order logic and undirected probabilistic graphical models. A Markov logic network (MLN) is a set of first-order formulae with weights. Intuitively, the more evidence we have that a formula is true the higher the weight of this formula.

A set of ground atoms is a possible world. We say that a possible world W satisfies a ground formula F , and write $W \models F$, if F evaluates to true in W . The probability of a possible world W is given by

$$p(W) = \frac{1}{Z} \exp \left(\sum_{(F_i, w_i)} \sum_{G \in \mathcal{G}_{F_i}^C: W \models G} w_i \right).$$

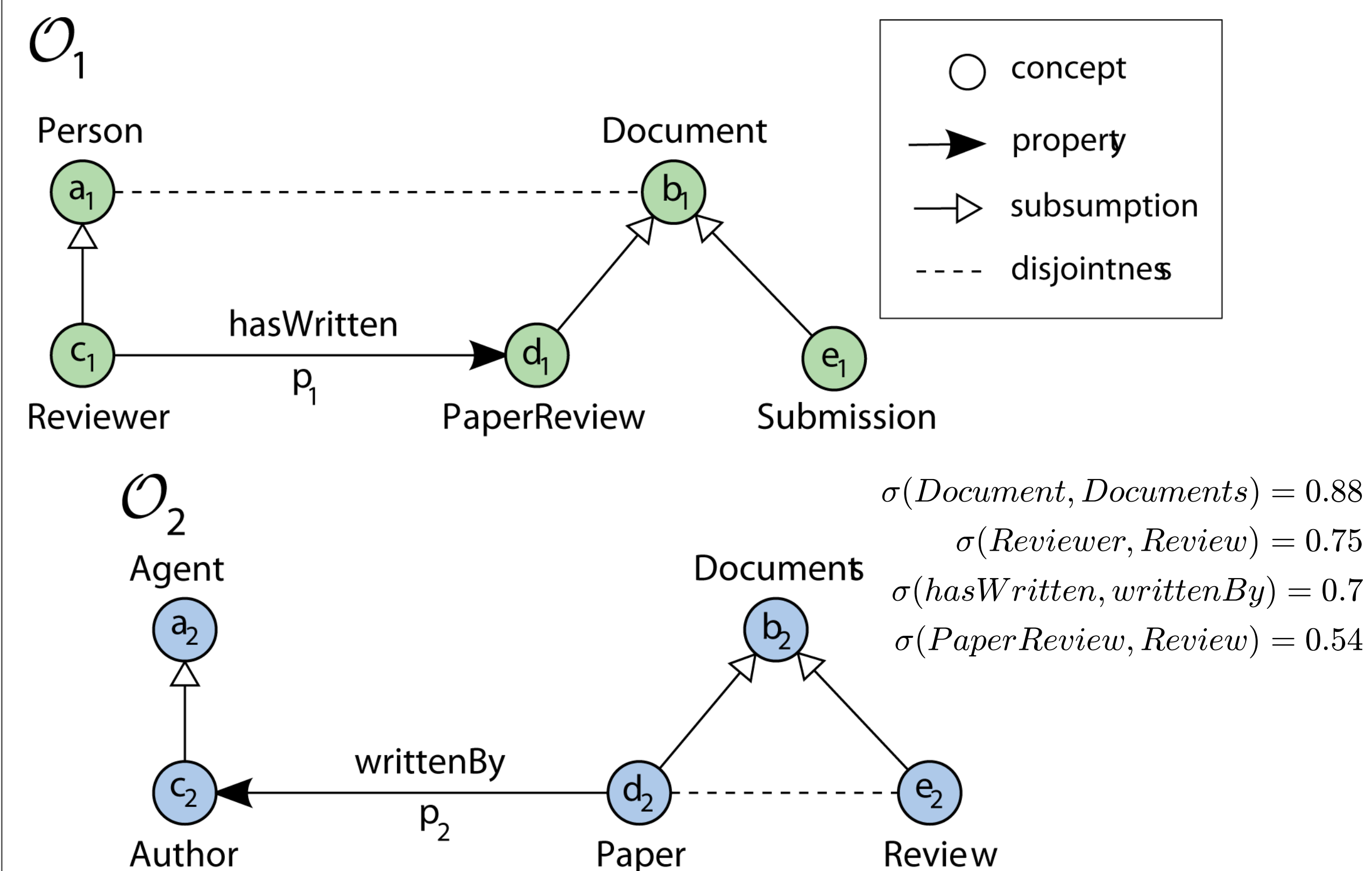
Markov Logic and Ontology Matching

- We introduce **observable predicates** O to model the structure of \mathcal{O}_1 and \mathcal{O}_2 with respect to both concepts and properties

$$\begin{aligned} \mathcal{O}_i \models D \sqsubseteq E &\mapsto sub_i(d, e) \\ \mathcal{O}_i \models D \sqsubseteq \neg E &\mapsto dis_i(d, e) \\ \mathcal{O}_i \models \exists R. \top \sqsubseteq D &\mapsto sub_i^d(r, d) \\ \mathcal{O}_i \models \exists R. \top \sqsupseteq D &\mapsto sup_i^d(r, d) \\ \mathcal{O}_i \models \exists R. \top \sqsubseteq \neg D &\mapsto dis_i^d(r, d) \\ &\dots \end{aligned}$$

- Knowledge encoded in the ontologies is assumed to be true \rightarrow ground predicates of observable predicates are *hard constraints*
- Hidden predicates** m_c and m_p , on the other hand, model the sought-after concept and property correspondences, respectively
- Ground atoms of these hidden predicates are assigned the weights specified by the a-priori similarity measure σ
- The higher this value for a correspondence the more likely the correspondence is correct **a-priori**

MAP Inference as Alignment Process



A-priori confidence values

$$\begin{aligned} (m_c(c, d), \sigma(C, D)) &\quad \text{if } C \text{ and } D \text{ are concepts} \\ (m_p(p, r), \sigma(P, R)) &\quad \text{if } P \text{ and } R \text{ are properties} \end{aligned}$$

Cardinality Constraints

$$\begin{aligned} m_c(x, y) \wedge m_c(x, z) &\Rightarrow y = z \\ m_c(x, y) \wedge m_c(z, y) &\Rightarrow x = z \end{aligned}$$

Coherence Constraints

$$\begin{aligned} dis_1(x, x') \wedge sub_2(x, x') &\Rightarrow \neg(m_c(x, y) \wedge m_c(x', y')) \\ dis_1^d(x, x') \wedge sub_2^d(y, y') &\Rightarrow \neg(m_p(x, y) \wedge m_c(x', y')) \\ &\dots \end{aligned}$$

Stability Constraints

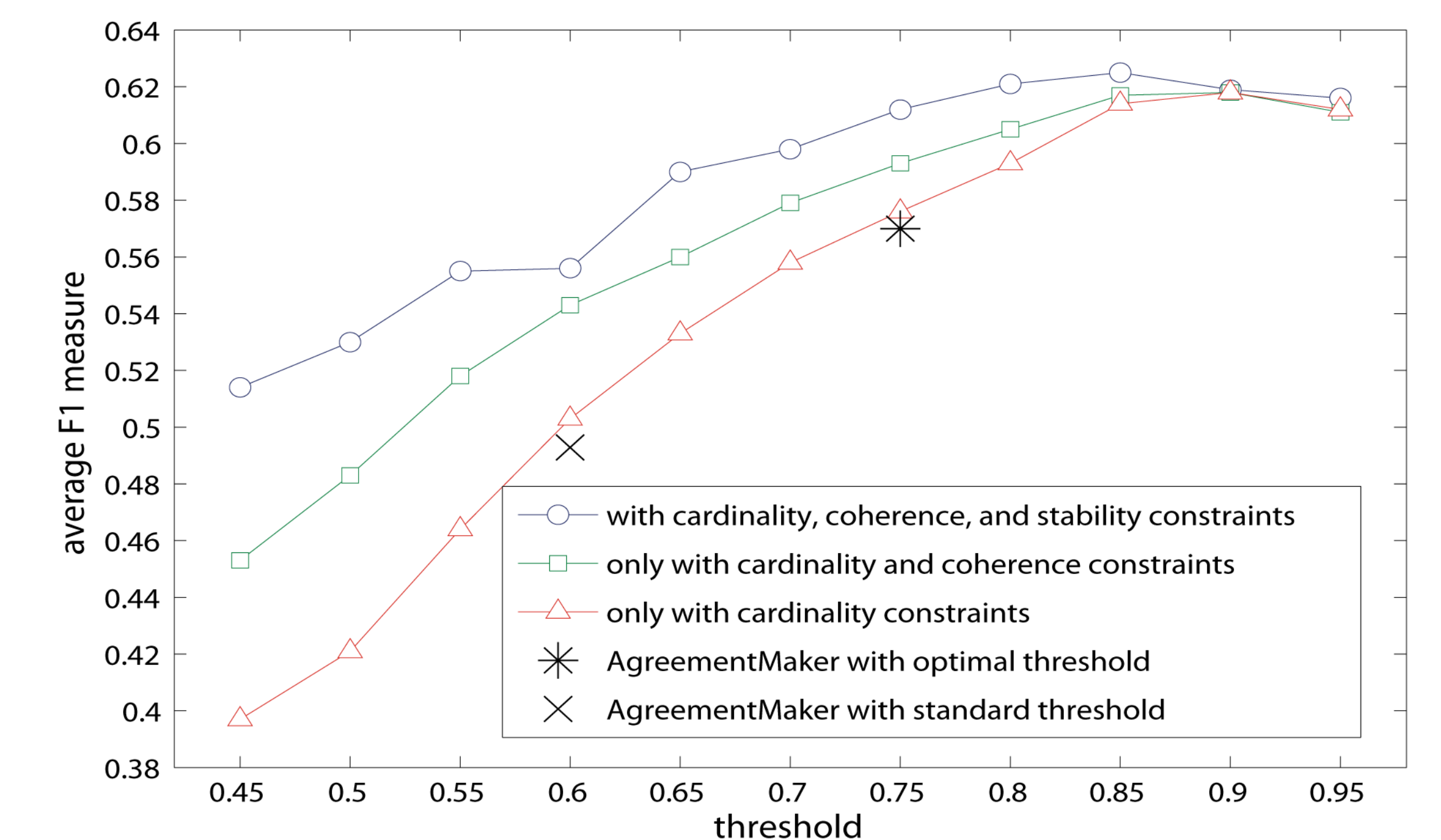
$$\begin{aligned} (sub_1(x, x') \wedge \neg sub_2(y, y')) &\Rightarrow m_c(x, y) \wedge m_c(x', y'), w_1 \\ (sub_1^d(x, x') \wedge \neg sub_2^d(y, y')) &\Rightarrow m_p(x, y) \wedge m_c(x', y'), w_2 \\ &\dots \end{aligned}$$

Maximize: $0.88x_1 + 0.75x_2 + 0.7x_3 + 0.54x_4$
Subject to:

$$\begin{aligned} x_2 + x_4 &\leq 1 \\ x_1 + x_3 &\leq 1 \\ x_1 + x_2 &\leq 1 \\ x_2 + x_3 &\leq 1 \end{aligned}$$

Experiments

- The Ontofarm dataset is basis for the experiments; it is the evaluation dataset for the OAEI conference track consisting of several ontologies modeling the domain of scientific conferences
- Reference alignments for 21 pairs of conference ontologies are made available by the organizers of the OAEI
- For the a-priori similarity σ we used a simple string similarity measure based on the Levenstein distance
- We applied the reasoner Pellet to create the ground MLN formulation
- TheBeast to convert the MLN formulations to the corresponding ILPs
- Mixed integer programming solver SCIP to solve the ILPs
- Four different ML formulations: Only cardinality constraints (**ca**); cardinality and coherence constraints (**ca+co**); cardinality, coherence, and stability constraints (**ca+co+sm** and **ca+co+sl**)



F₁ scores for **ca**, **ca+co**, and **ca+co+sm** averaged over the 21 OAEI reference alignments for thresholds ranging from 0.45 to 0.95. *AgreementMaker* was the best performing system on the conference dataset of the latest ontology evaluation initiative in 2009.

threshold	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
ca+co+sm	0.56	0.59	0.60	0.61	0.62	0.63	0.62	0.62
ca+co+sl	0.57	0.58	0.58	0.61	0.61	0.61	0.63	0.62
ca+co	0.54	0.56	0.58	0.59	0.61	0.62	0.62	0.61

Average F₁-values over the 21 OAEI reference alignments for manual weights (**ca+co+sm**) vs. learned weights (**ca+co+sl**) vs. formulation without stability constraints (**ca+co**); thresholds range from 0.6 to 0.95.

Future Work

The framework is not only useful for aligning concepts and properties but can also include instance matching. For this purpose, one would only need to add a hidden predicate modeling instance correspondences. The resulting matching approach would immediately benefit from probabilistic joint inference, taking into account the interdependencies between terminological and instance correspondences.