Semantical evaluators

Antoon Bronselaer[1] Guy De Tré[1]
1. Department of Telecommunications and Information Processing, Ghent University
Ghent, Belgium
Email: {antoon.bronselaer,guy.detre}@ugent.be

Abstract— *In the context of a possibilistic framework for detection of object co-reference, evaluators have been defined as operators that compare two values and express the belief that such values are co-referent. Hereby, co-reference of two values means that these values describe the same entity in the real world. In this paper, a class of evaluators is investigated that determines the belief of co-reference based on semantical connections between values of the universe. These semantical connections are modeled by means of binary relations. In case these binary relations are not a-priori given, they can be (partially) learned in an iterative co-reference detection schema.*

Keywords— Possibility theory, co-reference, binary relations

Table 1: Sample of ‘restaurant’ dataset

name	address	city	type
campanile	624 s. la brea ave.	los angeles	american
campanile	624 s. la brea ave.	los angeles	californian
grill on the alley	9560 dayton way	los angeles	american
grill the	9560 dayton way	beverly hills	american

1 Introduction

In order to deal with the increasing amount of (partially) independent information sources, detection of duplicate information storage is of key importance to avoid inefficient storage management and inconsistencies. In the context of this paper, data are discussed in terms of *objects*. An object is a structured description of a real world entity. Hereby, structured refers to a complex user defined data structure that is used to model data. Such is the case with, among others, relational databases, XML-documents and Object Oriented environments. Objects that refer to the same entity are called *co-referent* objects. A typical property of an object is that it can be decomposed into sub-objects with a *well-defined* domain. These sub-objects are called *attributes*. Note that attributes are not bound to have an atomic data structure, i.e. any collection type is a valid datatype for an attribute. In case of relational databases, the attributes are the tuple fields.

As a running example throughout this paper, the ‘restaurant’ dataset [1] is used, which is a table from a relational database that describes restaurants. Each restaurant is represented by an object that has four attributes: name, address, city and type. Next to the attributes, the dataset contains a source field, which indicates the restaurant guide from which the information was taken (‘fodor’ or ‘zagat’) and an identifier field ‘oid’. The goal of this dataset is to find the co-referent restaurants. The dataset is chosen as an example in this paper because syntactical evaluators are not suited for attributes ‘city’ and ‘type’. It is shown in this paper how semantical evaluators can be used for these attributes. A sample of the dataset is shown in Table 1, where objects 1 and 3 have source ‘fodor’ and objects 2 and 4 have source ‘zagat’. Objects 1 and 2 are co-referent objects and so are objects 3 and 4.

In a possibilistic setting, an evaluator E_U determines the possibility (or belief) of (non) co-reference about two values from a universe U. This universe can be the domain of complex objects or the domain of less complex attributes. The subject of this paper is to investigate a special class of evaluators called *semantical evaluators*. These evaluators use binary relations on the universe of interest to determine the belief that two values are co-referent. The transfer of knowledge from a relational level to a comparative level has already been studied in literature. The first notion of semantical connections between linguistic terms is pointed out by Quillian [2]. Resnik uses a network of semantical connections to construct a *semantical similarity measure* [3]. The key idea in this model is to compare the information that is held by two concepts in order to compute the similarity between two words. Further on, Resnik uses an existing taxonomy as knowledge base. This interesting approach has led to several further enhancements, which are all based on information theory [4, 5]. The knowledge base T reflects relations between elements of a universe of concepts. Each approach then uses T to construct a similarity measure. The research presented in this paper is in line of these approaches and provides a theoretical framework for semantical evaluators. The case in which the the taxonomy T is not a-priori known, will be treated in this paper, leading to an unsupervised schema for construction of binary relations. Unsupervised approaches for co-reference have been studied in [6, 7, 8]. The idea proposed here is in fact complementary with the key idea behind these approaches. In recent research, the term co-reference has become relevant in the field of computational linguistics [9, 10, 11]. The search for co-referent *non-structured* objects, i.e. flat text documents, has important applications in text clustering and information aggregation. In this recent field, semantics is of crucial importance to obtain good cluster accuracy. However, it is emphasized that semantical evaluators as introduced in this paper operate in a framework for *structured* objects, in contradiction to the frameworks for *non-structured* objects in [9, 10, 11]. The remainder of the paper is structured as follows. Section 2 summarizes some basic notations and preliminary knowledge. In Section 3, semantical evaluators are defined and properties are investigated. Section 4 introduces the iterative co-reference

ISBN: 978-989-95079-6-8

detection scheme, which is an unsupervised learning schema in which binary relations used by semantical evaluators are increasingly constructed. Finally, Section 5 summarizes the main contributions of this paper.

2 Preliminaries

Within the context of this paper, a (structured) object is a (structured) description of a real world entity that is decomposable into *attributes* and *co-referent* objects are objects that describe the same entity. The possibilistic model for detection of co-referent objects is based on the concept of *possibilistic truth values* [12, 13, 14]. A possibilistic truth value $\tilde{p}$ is a normalized possibility distribution over the domain $\mathbb{B} = \{T, F\}$ and represents uncertainty about the truth value of a proposition p. Hence $\tilde{p} = \{(T, \mu_{\tilde{p}}(T)), (F, \mu_{\tilde{p}}(F))\}$. A short notation thereof is $\tilde{p} = (\mu_{\tilde{p}}(T), \mu_{\tilde{p}}(F))$. The domain of all possibilistic truth values is denoted $\tilde{\wp}(\mathbb{B})$ which is complete lattice. Total order relations $\tilde{R}$ on this domain are defined as follows, with R a total order relation on the unit interval $[0, 1]$:

$$\tilde{p}_1 \; \tilde{R} \; \tilde{p}_2 \Leftrightarrow \begin{cases} \mu_{\tilde{p}_2}(F) \; R \; \mu_{\tilde{p}_1}(F), & \mu_{\tilde{p}_1}(T) = \mu_{\tilde{p}_2}(T) = 1 \\ \mu_{\tilde{p}_1}(T) \; R \; \mu_{\tilde{p}_2}(T), & \text{otherwise} \end{cases}$$

Possibilistic truth values describe available knowledge (or belief) about a proposition. It is emphasized that they do not describe a state of graded truth. The most important operators for knowledge aggregation in the domain of possibilistic truth values are generalized *conjunction*

$$\tilde{\wedge} : \tilde{\wp}(I)^2 \rightarrow \tilde{\wp}(I) : \tilde{p}\tilde{\wedge}\tilde{q} \mapsto \{ \, (T, \min\left(\mu_{\tilde{p}}(T), \mu_{\tilde{q}}(T)\right)) \, , \\ (F, \max(\mu_{\tilde{p}}(F), \mu_{\tilde{q}}(F))) \, \}$$

and *disjunction*

$$\tilde{\vee} : \tilde{\wp}(I)^2 \rightarrow \tilde{\wp}(I) : \tilde{p}\tilde{\vee}\tilde{q} \mapsto \{ \, (T, \max(\mu_{\tilde{p}}(T), \mu_{\tilde{q}}(T))) \, , \\ (F, \min(\mu_{\tilde{p}}(F), \mu_{\tilde{q}}(F))) \, \}$$

In the context of co-reference detection, possibilistic truth values are generated by *evaluators*. These are formally defined as follows:

Definition 1 (Evaluator)
Assume a universe U of values describing entities in the real world. An evaluator on U is a commutative function E_U such that:

$$E_U : U^2 \rightarrow \tilde{\wp}(\mathbb{B}) : (u, v) \mapsto E_U(u, v) = \tilde{p} \qquad (1)$$

with p =“u and v are co-referent”, $\mu_{\tilde{p}}(T)$ is the possibility that p is true and $\mu_{\tilde{p}}(F)$ is the possibility that p is false.

An evaluator is *reflexive* if $\forall (u, u') \in U^2 : u = u' \Rightarrow E_U(u, u') = \{(T, 1)\}$ and is *strong reflexive* if $\forall (u, u') \in U^2 : u = u' \Leftrightarrow E_U(u, u') = \{(T, 1)\}$.

This paper deals with evaluators that extract knowledge from binary relations. A binary relation R on U is a subset of the Cartesian product $U \times U$. The elements of R are thus couples of elements from U. $(u, v) \in R$ is also denoted as $u \; R \; v$. An equivalence relation is a binary relation that is *reflexive* ($\forall u \in U : u \; R \; u$), *symmetrical* ($\forall (u, v) \in U^2 : u \; R \; v \Leftrightarrow v \; R \; u$) and *transitive* ($\forall (u, v, w) \in U^3 : (u \, R \, v \wedge v \; R \; w) \Rightarrow u \; R \; w$). A partial order relation is a binary relation that is reflexive, transitive and anti-symmetrical ($\forall (u, v) \in U^2 : (u \, R \, v) \wedge (v \; R \; u) \Rightarrow (u = v)$). A *strict* partial order relation is anti-reflexive ($\forall u \in U : \neg(u \; R \; u)$), anti-symmetric and transitive.

3 Semantical evaluators

3.1 Definitions

Evaluators have been introduced and investigated in [15]. In general, three classes of evaluators are distinguished: syntactical, semantical and hybrid evaluators. *Syntactical* evaluators determine possibility based on syntactical differences and are useful to deal with typographical errors in string data and inaccuracies with numerical data. *Semantical* evaluators are based on *binary relations* between elements of a universe U. The name ‘semantical evaluator’ reflects that in many cases, such binary relations model a semantical connection between elements of the universe. However, the use of the framework presented here is not restricted to semantical connections only. In fact, binary relations can very well express a syntactical connection, rather than a semantical one, for example the *equality* relation. Nevertheless, the applications of interest assume in most cases relations that are semantical in nature, which explains the name of this type of evaluators. *Hybrid* evaluators are a combination of several evaluators on the same domain. Syntactical evaluators are studied in [15], whereas semantical evaluators are the subject of this paper.

The rationale of a semantical evaluator emerges from the fact that co-reference of objects is nothing more than equality of the entities described by the objects. As such it can be seen that co-reference of objects can in fact be modeled by a symmetrical reflexive binary relation (say R_c for convenience) in the domain $O \times O$ with O the set of objects. Moreover, the domain O can be decomposed into sub-domains U_i which are the domains of attributes a_i. Hence, if a binary relation R_i is defined in $U_i \times U_i$, a key question is then how the information given by relation R_i can be translated into information about R_c. It is thus studied in this paper how a given binary relation R on a universe U can be used to generate the possibility that elements of U belong to co-referent objects. An important assumption hereby made is that R contains *positive* knowledge about co-reference, which means that connection of two elements u and v through a binary relation is regarded as evidence for their co-reference. If elements are not connected through R, it is concluded that u and v are not co-referent. This is formalized as follows.

Definition 2 (Semantical evaluator)
Assume a universe U and a binary relation R. A semantical evaluator $E_{U,R}$ is defined by means of a function f_R called the knowledge transfer function, such that $\forall (u, v) \in U^2$: $E_{U,R}(u, v) = f_R(u, v)$ with:

$$f_R : U^2 \rightarrow \tilde{\wp}(\mathbb{B}) \qquad (2)$$

such that

$$f_R(u, v) = \begin{cases} \tilde{p}_{u,v} & u \; R \; v \vee v \; R \; u \\ (0, 1) & \neg(u \; R \; v) \wedge \neg(v \; R \; u) \end{cases} \qquad (3)$$

The condition $u \; R \; v \vee v \; R \; u$ in Definition 2 is a consequence of the assumption of commutativity in the definition of evaluators. As an example, let $\mathcal{S}$ be the domain of strings and consider the relation R=‘city-part-of’. If, for the sake of simplicity, it is assumed that $\tilde{p}_{u,v}$ is independent of u and v and equal to $(1, 0.3)$, then:

$$E_{\mathcal{S},R}(\text{“Manhattan”}, \text{“New York”}) = (1, 0.3)$$
$$E_{\mathcal{S},R}(\text{“New York”}, \text{“Chicago”}) = (0, 1)$$

The assignment of possibilities by a semantical evaluator is constrained by the specificity of the relation R. Assume two binary relations R_1 and R_2 on universe U such that $R_1 \subset R_2$, then R_2 applies for more couples of elements than does R_1. In an extreme case, a relation R can apply for each couple of elements, which means that R is an equivalence relation on U with 1 equivalence class. Clearly such a relation offers no information about co-reference at all, which means that:

$$R = U^2 \Rightarrow \forall (u,v) \in U^2 : E_R(u,v) = (1,1) \qquad (4)$$

Thus, if more couples satisfy a relation R, R is less specific and less knowledge about co-reference can be derived from R. Hence, a knowledge transfer function must satisfy:

$$(R_1 \subseteq R_2) \Rightarrow \forall (u,v) \in R_1 : f_{R_1}(u,v) \tilde{\geq} f_{R_2}(u,v) \qquad (5)$$

The principle of specificity can also be applied inside a relation R. Consider a binary relation R defined over the universe U. The projection of R over an element $u \in U$ is given by:

$$\text{Proj}_u(R) = \{(x,y) | x\ R\ y \wedge (x = u \vee y = u)\} \qquad (6)$$

The knowledge transfer from R to $\tilde{\wp}(\mathbb{B})$ modeled by f_R is *monotone* if and only if for any triple $(u,v,w) \in U^3$ that satisfies $uRw \vee wRu$ and $vRw \vee wRv$, we have that:

$$|\text{Proj}_u(R)| \geq |\text{Proj}_v(R)| \Leftrightarrow f_R(u,w) \tilde{\leq} f_R(v,w) \qquad (7)$$

It can be verified that monotone knowledge transfer is obtained if f_R is constructed as follows:

$$\begin{aligned} & u\ R\ v \vee v\ R\ u \Rightarrow \\ & f_R(u,v) = \left(1, g\left(\tfrac{|\text{Proj}_u(R)|}{2|R|-1}, \tfrac{|\text{Proj}_v(R)|}{2|R|-1}\right)\right) \end{aligned} \qquad (8)$$

with g a monotonic increasing, commutative function such that $g(1,1) = 1$. Monotone transfer knowledge implies that it is *more* certain that u and v are co-referent if they are connected by R and if there is *less* evidence that u or v are co-referent with other elements.

A common problem with semantical evaluators is that relations are a-priori information. In other words, relations are constructed without observing the actual values on which the evaluator must operate. It is possible that the values on which R is constructed are not *equal* to values of an observed problem, but rather co-referent. This is caused by the fact that values considered for a-priori construction of R are typically standardized, which is not the case for values that are the subject of co-reference detection. For example, with R='city-part-of', it is known that "Manhattan" R "New York". Nevertheless:

$$E_{\mathcal{S},R}(\text{“Manhatan”}, \text{“New York”}) = (0,1)$$

due to the spelling error in "Manhatan". To deal with such a situation, an evaluator can be used to detect co-references between values from an observed problem and values from the a-priori universe on which R is constructed. Typically, such an evaluator will be *syntactical*. A semantical evaluator is then used in a *chain* of evaluators.

$$\begin{tikzcd} u \arrow[r, dashed, "E'_U(u{,}u')"] \arrow[d, leftrightarrow, "E_U(u{,}v)"'] & u' \arrow[d, leftrightarrow, "E_{U,R}(u'{,}v')"] \\ v \arrow[r, dashed, "E'_U(v{,}v')"'] & v' \end{tikzcd}$$

Figure 1: Evaluator chain

Definition 3 (Evaluator chain)
Assume a universe U equipped with a semantical evaluator $E_{U,R}$ and a reflexive evaluator E'_U. An evaluator chain is an evaluator E_U such that:

$$E_U(u,v) = \max_{(u',v')} E'_U(u,u') \tilde{\wedge} E_{U,R}(u',v') \tilde{\wedge} E'_U(v',v) \qquad (9)$$

The principle of chaining is schematically depicted in Figure 1. Due to reflexivity of E'_U it can be seen that:

$$(u\ R\ v \vee v\ R\ u) \Rightarrow E_U(u,v) = E'_U(u,v) \qquad (10)$$

Moreover, if E'_U is strong reflexive:

$$(u\ R\ v \vee v\ R\ u) \Leftrightarrow E_U(u,v) = E'_U(u,v) \qquad (11)$$

In the context of iterative co-reference detection (Section 4), an important concept is that of a hybrid evaluator.

Definition 4 (Hybrid evaluator)
Assume a universe U and a set of evaluators $\mathcal{E} = \{E_{U,1}, ..., E_{U,n}\}$ on U. A hybrid evaluator is an evaluator on U such that:

$$E_U = c(E_{U,1}, ..., E_{U,n}) \qquad (12)$$

with c a commutative, associative and monotonic increasing function called the combinatory function.

3.2 Examples

The most obvious example of a semantical evaluator is an evaluator that is based on the equality relation. In many cases, reflexivity of an evaluator is required which means that $\forall (u,v) \in U^2 : u = v \Rightarrow E_U(u,v) = (1,0)$. Hence, in order to support this constraint, equality of elements must imply full impossibility of non co-reference. As mentioned in Section 3.1, this evaluator can be regarded as syntactical evaluator, because the binary relation is not semantical in nature. A related case is a semantical evaluator where R is relaxed to an *equivalence* relation, which has interesting applications such as detection of synonyms, i.e. words with the same meaning. Another example is the case where two words don't have the same meaning, but they both fully qualify for description of a given entity. This is because the entity can not be described exactly and more descriptions are possible. In that case, words are equivalent with respect to their appropriateness for description of an entity, despite the fact that these words do not have the same meaning. This is the case with the attribute 'type' of the 'restaurant' dataset, where for example 'steakhouse' and 'american' are possible (and thus equivalent) descriptions of the type of restaurant.

Assume a universe U and assume a partition of U into n disjunct subsets U_i. Consider $n-1$ relations R_i such that each R_i is a surjective function between U_i and U_{i+1}. The set of relations $\{R_i\}_{i=1,2,\ldots,n-1}$ is called a *hierarchy* on U. It can be seen that, by definition, each R_i from a hierarchy is anti-reflexive and anti-symmetrical. Due to the fact that $\forall i \in \{1,...,n-1\} : im(R_i) = dom(R_{i+1})$, it is possible to compose relations R_i and R_{i+1}.

$$\forall(u,w) \in U^2 : u\, R_1 R_2\, w \Leftrightarrow (\exists v \in U : u\, R_1\, v \wedge v\, R_2 w) \tag{13}$$

Hence, $R_1R_2...R_n$ denotes the n-ary composition of n relations and thus, elements of non subsequent partitions are related through composition. For each hierarchy $\{R_i\}_{i=1,2,\ldots,n-1}$, consider the set of relations $\{R_i^*\}$ such that:

$$u\, R_i^*\, v \Leftrightarrow \exists k \leq l \leq i : u\, R_k...R_l v \tag{14}$$

By definition:

$$\forall i,j \leq n : i < j \Rightarrow R_i^* \subset R_j^* \tag{15}$$

It is then possible to construct $n-1$ semantical evaluators with knowledge transfer function $f_{R_i^*}$. Due to the consistency constraint:

$$\forall i,j \leq n : i < j \Rightarrow f_{R_i^*} \tilde{\geq} f_{R_j^*} \tag{16}$$

It can be seen that composing all relations, leads to a strict partial order relation. Hence, each partial order relation can be decomposed into relations with higher granularity and thus higher belief of co-reference. This principle can be applied on the 'city' attribute of the restaurant dataset, because the values are connected by a strict partial order relation.

4 Iterative co-reference detection

4.1 Approach

In this Section, attention is given to the case where R is not a-priori given. This is the case when dealing with subjective information, for example the 'type' attribute of the restaurant dataset. There is no a-priori information about which types can be used to denote the same restaurant. In that case a semantical evaluator can be (partially) constructed based on a training set. However, in what follows a method is proposed to construct a semantical evaluator in cases where no training set is given. Assume objects consisting of n attributes a_i. For each attribute a_i, an evaluator $E_{dom(a_i)} = E_i$ must be constructed. The set of attributes can thus be partitioned into attributes that require a syntactical evaluator (A_{syn}) and attributes that require a semantical evaluator (A_{sem}), such that $A = A_{syn} \cup A_{sem}$ and $A_{syn} \cap A_{sem} = \emptyset$. For the 'restaurant' dataset, $A_{syn} = \{$'name', 'street'$\}$ and $A_{sem} = \{$'type', 'city'$\}$.

The key idea of iterative co-reference detection, is to find the co-referent objects in m iterations. Semantical evaluators for which no binary relation R is available are gradually constructed throughout the different iterations. An evaluator is called *unavailable* if it is a semantical evaluator for which the binary relation R is unknown. In each iteration j, a subset of the available evaluators is selected. With these evaluators, a candidate set C_j of co-referent object couples is generated. For each attribute a_i that requires a semantical evaluator, a binary relation R_j is constructed by projecting each object couple from C_j over the attribute a_i. In other words, if the object couple (o_1, o_2) is contained in the candidate set C_j and attribute a_i requires a semantical evaluator, then the couple $(\text{Proj}_{a_i}(o_1), \text{Proj}_{a_i}(o_2))$ is considered as a couple of co-referent values for attribute a_i. As a consequence, this couple of values is added to the binary relation R_j. For each attribute a_i that requires semantical evaluation, a novel relation is R_j is constructed in each iteration j. The reason to do this is explained next. Typically, the generation of the first candidate set is based on very stringent criteria. If this is not the case, many false positives are included in the candidate set, which implies that constructed binary relations contain a large amount of couples. As explained before, a binary relation with large cardinality implies inferred knowledge with much uncertainty (Eq. 4, Eq. 5). Throughout sequential iterations, the candidate set is expanded by (i) relaxation of syntactical evaluators and (ii) additional knowledge of semantical evaluators. As a consequence, there is a high probability that for $k \leq l$, $C_k \subset C_l$. If this is true, relations constructed in iteration k are subsets of the corresponding relations in iteration l. Due to the consistency constraint (Eq. 5), relations constructed in iteration l imply more uncertainty. Therefor, in order to preserve the level of (un)certainty in relations constructed in early iterations, each iteration constructs a new relation. Hence, for each attribute a_i that requires a semantical evaluator, m binary relations are constructed. Each of these relations $R_{i,j}$ yields a semantical evaluator $E_{i,R_{i,j}}$. The constructed evaluator for attribute a_i is a hybrid evaluator (Eq. 12) with $\tilde{\vee}$ as combinatory function:

$$E_i(u,v) = E_{i,R_{i,1}}(u,v)\tilde{\vee}...\tilde{\vee}E_{i,R_{i,m}}(u,v) \tag{17}$$

Hence, the constructed evaluator tries to find the most stringent binary relation by which the values are connected as this relation yields the most belief in co-reference. After iteration m, the last candidate set C_m of co-referent couples is generated which gives us the possible co-referent objects. It is noted that iterative co-reference detection is a stand-alone comparison process: it produces a set of possible co-referent couples. However, it can also be used to construct evaluators and their corresponding binary relations. In other words, the proposed method can be used to find co-referent objects, but it's purpose

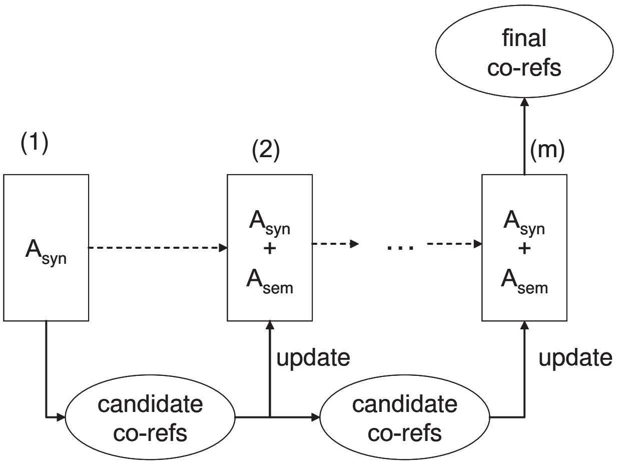

Figure 2: Iterative learning of semantical evaluators

Table 2: Statistics of restaurant test

'type' evaluator	Co-refs	Non Co-refs
Equality	22	5018
Equality + R	100	27811

can also be limited to construction of binary relations, which makes it compatible with other (possibilistic) frameworks for co-reference detection. Figure 2 shows a schematical representation of iterative co-reference detection.

4.2 Example

The idea of iterative co-reference detection is illustrated on the 'restaurant' dataset. The focuss in this example lies on illustrating the construction of binary relations. The dataset consists of two lists ($|L_1| = 533$ and $|L_2| = 331$). Among the 176423 couples of restaurants, 112 couples are co-referent. For convenience, only two attributes are initially considered: 'name' and 'type'. It is assumed that a syntactical evaluator is used for 'name' and a semantical evaluator must be constructed for 'type'. For the first iteration, consider an evaluator for attribute 'name' based on equality (i.e. values are co-referent if they are equal). Among all possible couples, 83 couples have an equal name, of which 82 are identified as co-referent, based on ground truth (i.e. equal oid's). Hence, candidate set C_1 contains 83 couples of restaurants. Projecting this candidate set over the 'type' attribute and filtering identical couples out, yields 40 couples of possible co-referent restaurant types. These couples determine the binary relation R_1. Table 2 learns how R_1 allows for assignment of belief to 100 co-referent couples (89.3% of all co-referent couples). However, 27811 non-coreferent couples are also assigned belief (15.8% of all non co-referent couples). This is partially due to the type of data. As can be seen, 5018 non-coreferent couples have equal values for type, which is an indication that the 'type' attribute is prone to many false positives. It must be remembered that the goal is to construct an evaluator that is used in a larger object comparison system. Hence, the contribution of semantical evaluators lies in the assignment of belief to co-referent values, thereby trying to minimize the belief attached to non co-referent values. As can be seen in Table 2, the fraction of non-coreferent values that are assigned belief can still be large in absolute numbers. However, this fraction can be filtered out using an intelligent aggregation function to combine the belief assigned by the different evaluators. The constructed relation for 'type' assigns belief to 89.3% of the co-referent couples, after one iteration. Assume in a second iteration, that the 'street' attribute is involved, which requires a syntactical evaluator. For the sake of simplicity, assume an evaluator for 'street' based on equality. If the evaluations for 'name' and 'street' are combined through disjunction, i.e. two restaurants are co-referent if their names are equal *or* their streets are equal, then the candidate set C_2 contains 106 object couples of which 101 are actually co-referent. Suppose that iteration 2 would be the last iteration, then the model can be refined by adding semantical evaluation of the 'type' attribute, i.e. two restaurants are co-referent if their types are connected by R_1 *and* (names are equal *or* their streets are equal). Using this model results in a candidate set of 95 couples of which 94 are co-referent, which means that 83.9% of all co-referent couples are detected with a precision of 98.95%. If iteration 2 is not the last iteration, then C_2 can be used to construct a second binary relation R_2. Note that by construction $R_1 \subset R_2$.

An advantage of iterative co-reference detection is that the binary relations R are dynamically adapted throughout the different iterations. If an approach with a training set is used, the relation R can be partially constructed only on the observations in the training set. With iterative co-reference detection, binary relations are constructed by generating a candidate set from *all* known observations. Hence, the relations constructed in an iterative co-reference detection scheme will be more complete.

5 Conclusion

In the possibilistic framework for detection of complex co-referent objects, a central concept is that of an evaluator, which is a function that estimates belief/uncertainty about co-reference of (sub)-objects. This paper introduces a formal framework for *semantical* evaluators E_U that model the transfer of knowledge from binary relations on the universe, to knowledge about co-reference of elements. This transfer is modeled by means of a knowledge transfer function f_R. It is shown that, depending on the application, such a function must satisfy consistency constraints. It is also shown how semantical evaluators can be used in an evaluator chain and how to combine evaluators in a hybrid evaluator. Finally, when relations on U are unknown, it is investigated how to (partially) construct such relations without use of a pre-labeled training set. This construction process is called iterative co-reference detection.

References

[1] Sheila Tejada, Craig Knoblock, and Steven Minton. Learning object identification rules for information integration. *Information Systems*, 26(8):607–633, 2001.

[2] Ross Quillian. *Semantic Information Processing*, chapter Semantic memory. MIT Press, 1968.

[3] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[4] Jay Jiang and David Conrath. Semantical similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*, pages 15–33, 1997.

[5] Yuhua Li. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, 2003.

[6] Matthew Jaro. Advances in record linking methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Society*, 84(406):414–420, 1989.

[7] William Winkler. Improved decision rules in the fellegi-sunter model of record linkage. Technical Report RR93/12, Statistical Research Report Series, 1993.

[8] William Winkler. Methods for record linkage and bayesian networks. Technical Report RRS2002/05, Statistical Research Report Series, 2002.

[9] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[10] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–111, Philadelphia, 2002.

[11] Xiaoqiang Luo and Imed Zitouni. Multi-lingual coreference resolution with syntactic features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 660 – 667, 2005.

[12] Lotfi Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[13] Henri Prade. Possibility sets, fuzzy sets and their relation to lukasiewicz logic. In *Proceedings of the International Symposium on Multiple-Valued Logic*, pages 223–227, 1982.

[14] Gert De Cooman. Towards a possibilistic logic. in: Fuzzy set theory and advanced mathematical applications, edited by da ruan, kluwer academic, pp. 89–133, boston., 1995.

[15] Antoon Bronselaer, Axel Hallez, and Guy De Tré. Evaluation in the possibilistic framework for object matching. In *Proceedings of the IPMU 08 Congres*, pages 1701–1708, Malaga, 2008.