Fuzzy Constraint Based Answer Validation

Steven Schockaert, Martine De Cock, and Etienne E. Kerre

Ghent University, Department of Applied Mathematics and Computer Science,
Fuzziness and Uncertainty Research Unit,
Krijgslaan 281 - S9, B-9000 Gent, Belgium
{Steven.Schockaert, Martine.DeCock, Etienne.Kerre}@ugent.be

Abstract. Answer validation is an important component of any question answering system. In this paper we show how the formalism of prioritized fuzzy constraint satisfaction allows to unify and generalize some common validation strategies. Moreover, answer candidates are represented by fuzzy sets, which allows to handle imprecise answers.

1 Introduction

Question answering systems try to improve the functionality of search engines by providing an exact answer to a user's question, rather than a list of documents. A typical question answering system consists of a question analysis module, a search engine, an answer extraction module and an answer validation module. At least two fundamentally different ways to handle answer validation are used by current systems. Corpus–based methods (e.g. [5]) rely on a deep linguistic analysis of the question and the answer candidates, while redundancy–based methods (e.g. [2],[3],[6]) rely on the massive amount of information available on the web. This paper will focus on the latter kind of methods.

Since it is reasonable to assume that on the web, the answer to most questions is stated in a lot of documents, we can assume that there will be documents in which the answer is formulated in a simple way. As a consequence, simple answer extraction algorithms often suffice. However, simplicity comes with a price; a lot of web pages contain incorrect information, so the answer validation process used in corpus–based methods is not appropriate. Most redundancy–based methods apply some kind of voting: the answer which occurs most often is considered the most likely answer to be correct. This approach has the disadvantage of favouring short, unspecific, answers (e.g. "1928" over "July 26, 1928"). Some systems (e.g. [2],[6]) therefore apply heuristics to boost the scores of specific answers. These heuristics would treat an occurrence of "1928" as evidence for "July 26, 1928" which, in our opinion, is not a fully satisfactory approach.

In this paper we propose an alternative voting scheme, which separates positive and negative information about the feasibility of the answer candidates. To this end, we represent answer candidates as fuzzy sets and define a degree of inconsistency and a degree of inclusion between answer candidates. We show how this scheme can be further refined by asking additional questions and enforcing fuzzy constraints on the results.

P.S. Szczepaniak et al. (Eds.): AWIC 2005, LNAI 3528, pp. 394–400, 2005.

2 Answer Comparison

Let's consider the question "When was the Mona Lisa painted?". When examining the first few snippets returned by Google[1] for this question, we find answers like "the 1500s", "1506", "1503–1506", ... It is clear that simple string equality won't yield very good results in this case. Instead we will represent each answer as a fuzzy set in a suitable universe which enables us to handle differences in granularity (e.g. "July 26, 1928" vs. "1928"), intervals (e.g. "1503–1506", "the 1920s") and vague descriptions (e.g. "the late 1920s", "around 1930").

Recall that a fuzzy set A on a universe U is a mapping from U to the unit interval $[0,1]$. If $A(u) = 1$ for some u in U, A is called normalised. To generalize the logical conjunction to the unit interval $[0,1]$, we have a large class of $[0,1]^2 - [0,1]$ mappings, called t-norms at our disposal. Likewise, logical implication can be generalized by a class of $[0,1]^2 - [0,1]$ mappings called implicators. For further details on t-norms and implicators we refer to [8].

Let a_1 and a_2 be two fuzzy sets in the universe $\mathcal{D}$ of dates, the degree of inclusion $incl(a_1, a_2)$ and the degree of contradiction $contr(a_1, a_2)$ between a_1 and a_2 can be given by

$$incl(a_1, a_2) = \inf_{u \in \mathcal{D}} I(a_1(u), a_2(u)) \qquad contr(a_1, a_2) = 1 - \sup_{u \in \mathcal{D}} T(a_1(u), a_2(u))$$

where I is an implicator and T is a t-norm. In our implementation we used the Łukasiewicz implicator I_W defined by $I_W(x, y) = \min(1, 1 - x + y)$ and the t-norm T_M defined by $T_M(x, y) = \min(x, y)$ for x and y in $[0,1]$. For each answer candidate a we can define the degree $pos(a)$ to which this answer is confirmed by the other candidates and the degree $neg(a)$ to which this answer is inconsistent with the other candidates:

$$pos(a) = \frac{1}{n} \sum_{i=1}^{n} incl(a_i, a) \qquad neg(a) = \frac{1}{n} \sum_{i=1}^{n} contr(a_i, a)$$

where $(a_1, a_2, \ldots, a_n)$ is the list of all answer candidates ($a_i = a_j$ may hold for some $i \neq j$, i.e. an answer candidate can occur several times). We interpret $pos(a)$ as the degree of feasibility of an answer candidate and $neg(a)$ as the degree of inconsistency, where $pos(a) = 1 - neg(a)$ doesn't hold in general. If I is a border implicator and all answer candidates are normalised then $pos(a) + neg(a) \leq 1$ holds. As a consequence, the set of answer candidates can be represented by an intuitionistic fuzzy set [1].

3 Refining the Answer Scores

Asking additional questions Prager et al. [9] introduced the idea to (automatically) ask additional questions in order to estimate the feasibility of an answer

[1] http://www.google.com

candidate. To answer the question "When did Leonardo da Vinci paint the Mona Lisa?", Prager et al. suggest to ask the additional questions "When was Leonardo da Vinci born?" and "When did Leonardo da Vinci die", which gives us the variables X_{work}, X_{born} and X_{died}. The possible instantiations of these variables are the answer candidates of the corresponding questions. All answer triplets that do not satisfy the following constraints[2] are rejected in [9]:

$$X_{born} + 7 \leq X_{died} \leq X_{born} + 100 \tag{1}$$

$$X_{work} \leq X_{died} \leq X_{work} + 100 \tag{2}$$

$$X_{born} + 7 \leq X_{work} \leq X_{born} + 100 \tag{3}$$

The use of crisp constraints has the disadvantage that a lot of world knowledge can not be expressed. For example, by using this kind of rather arbitrary threshold values, we can not express that it is more likely that someone became 70 years old than that someone became 8 years old. Another problem with this approach is how to combine the frequency counts of the answer candidates of the three variables X_{work}, X_{born} and X_{died}. In this section we show how both problems can be solved by using prioritized fuzzy constraints.

Prioritized fuzzy constraint satisfaction Let X_1, X_2, ..., X_n be variables taking values in the finite domains D_1, D_2, ..., D_n respectively. A fuzzy constraint c is a mapping from $D_1 \times D_2 \times \ldots \times D_n$ to the unit interval $[0, 1]$. For a constraint c and an instantiation $(x_1, x_2, \ldots, x_n) \in D_1 \times D_2 \times \ldots \times D_n$ of the variables, $c(x_1, x_2, \ldots, x_n)$ is interpreted as the degree to which the constraint c is satisfied by this instantiation. In [4] the notion of prioritized fuzzy constraint is introduced by assigning a priority to each constraint, which can be interpreted as the degree of importance of the constraint. Let α_i in $[0, 1]$ be the priority of constraint c_i ($i \in \{1, 2, \ldots, m\}$), the degree of joint satisfaction of the constraints $c_1, c_2, \ldots, c_m$ by an instantiation $(x_1, x_2, \ldots, x_n)$ can then be defined by [7]:

$$C(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{m} P(\alpha_i, c_i(x_1, x_2, \ldots, x_n)) \tag{4}$$

where P is $[0, 1]^2 - [0, 1]$ mapping called a priority operator. It is easy to see that the notion of a priority operator as defined in [7] corresponds to that of a border implicator.

Constructing the fuzzy constraints As a fuzzification of inequality (1) for example, we used the fuzzy constraint c_1 defined by

$$c_1(x_b, x_d) = incl(x_d \ominus_T x_b, f) \tag{5}$$

where (x_b, x_d) is an instantiation of (X_{born}, X_{died}); x_b and x_d are fuzzy sets in the universe of dates $\mathcal{D}$. According to the extension principle of Zadeh [10], $x_d \ominus x_b$ is the fuzzy set in the universe of real numbers $\mathbb{R}$ defined for d in $\mathbb{R}$ by

[2] Prager et al. [9] consider only crisp answer candidates, corresponding to a year.

$$(x_d \ominus_T x_b)(d) = \sup_{d_1 - d_2 = d} T(x_d(d_1), x_b(d_2)) \tag{6}$$

where T is a t-norm. The result of the date subtraction $d_1 - d_2$ is treated as a real number respresenting the number of years between the date d_1 and the date d_2. The fuzzy set f in the universe $\mathbb{R}$ reflects life expectation expressed in years and is defined for d in $\mathbb{R}$ by

$$f(d) = \begin{cases} \frac{d}{30} & \text{if } 0 \leq d \leq 30 \\ 1 & \text{if } 30 \leq d \leq 90 \\ \frac{120-d}{30} & \text{if } 90 \leq d \leq 120 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Fuzzification of (2) and (3) can be treated analogously. The priority of each of these fuzzy constraints is 1.

For a variable X, corresponding to a question with answer candidates x_1, $x_2, \ldots, x_n$, we can impose the unary constraint c_X, defined for each answer candidate x by[3]

$$c_X(x) = \frac{1 - neg(x)}{1 - neg(a^*)} \tag{8}$$

where $neg(a^*) = \inf_{a \in \mathcal{A}} neg(a)$ and $\mathcal{A}$ is the set of all answer candidates. The priority of this constraint can be interpreted as the reliability of the frequency count. In other words, if the number of answer candidates is high (resp. low) the priority should be high (resp. low) too. A possible definition of the priority α_X of the constraint c_X is given by $\alpha_X = \frac{n}{n+K}$ where n is the number of (not necessarily distinct) answer candidates as before, and $K > 0$ is a constant.

Yet another type of fuzzy constraints that can be imposed is based on co-occurrence. Consider the question "When was the Mona Lisa painted". In this case there is a fourth variable X_{pers} representing the painter of the Mona Lisa. If x_w is an answer candidate for the date that some person x_p painted the Mona Lisa, then we can assume that a lot of the sentences containing a date that is entailed by x_w in a set of documents about the "Mona Lisa" should contain a reference to x_p. We can express this by enforcing the constraint c_{assoc}, defined by

$$c_{assoc}(x_w, x_p) = \frac{assoc(x_w, x_p)}{\sup\limits_{(x^W, x^P)} assoc(x^W, x^P)} \tag{9}$$

where the supremum in (9) is taken over all possible instantiations (x^W, x^P) of (X_W, X_P) and $assoc(x_w, x_p)$ measures the extent to which sentences containing a date that is entailed by x_w tend to contain a reference to x_p.

[3] We will consider only constraints that are normalised (i.e. constraints for which there exists at least one possible instantiation that fully satisfies the constraint).

Putting the pieces together Let $\{c_1, c_2, \ldots, c_m\}$ be the set of all considered constraints and let X_1, X_2,..., X_n be the variables that are considered relevant to the user's question. The degree $neg_C(x_1, x_2, \ldots, x_n)$ of infeasibility of an answer tuple $(x_1, x_2, \ldots, x_n)$ is then given by

$$neg_C(x_1, x_2, \ldots, x_n) = 1 - C(x_1, x_2, \ldots, x_n) \tag{10}$$

where C is defined as in (4). For notational simplicity we use $c(x_1, x_2, \ldots, x_n)$ even when the constraint c doesn't refer to all x_i $(1 \leq i \leq n)$. The degree of feasibility $pos_C(x_1, x_2, \ldots, x_n)$ of an answer tuple $(x_1, x_2, \ldots, x_n)$ is defined by

$$pos_C(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} pos(x_i) \tag{11}$$

4 Experimental Results

To implement the ideas presented in this paper we extracted answer candidates using a simple pattern matching algorithm. Given the title of some work of art, possible creation dates and possible creators along with their birthdate and death date, are extracted from the snippets returned by Google for some (automatically generated) queries. Generic patterns to extract the entities of interest were constructed by hand. Table 1 shows some of the creators and creation dates that are found for the "Mona Lisa" together with their frequency of occurrence. Simply counting the frequency of occurrence of each answer candidate gives good results for determining the creator in this example; the creation date however is more problematic. In fact there exist several opinions about when the "Mona Lisa" was painted, but most agree it must have been between 1503 and 1506. For each potential creator our algorithm tries to discover the date this person was born and the date this person died. Using this information $pos_C(x_w, x_b, x_d, x_p)$ and $neg_C(x_w, x_b, x_d, x_p)$ are calculated for each instantiation (x_w, x_b, x_d, x_p) of the variables X_{work}, X_{born}, X_{died} and X_{pers}. The answer tuples $\overline{x}$ (i.e. the instantiations of the variables) are ranked using the product of $pos_C(\overline{x})$ and $1 - neg_C(\overline{x})$; the results are shown in table 2. We omit answer tuples that are entailed by another answer tuple that is ranked higher.

Table 1. Frequency counts for the creator and creation date of the "Mona Lisa"

Creator	Frequency
Leonardo da Vinci	18
Leonardo	8
Slick Rick	6
Everybody	6
Leonardo Da Vinci	6
Nick Pretzlik	2
Fernando Botero	2

Creation date	Frequency
1506	6
1950	5
1503	2
between 1503 and 1506	2
early 1500s	1
between 1503 and 1507	1
1502	1

Table 2. Top candidates of our algorithm for the "Mona Lisa"

Creator	Creation date	Score
Leonardo da Vinci (1452 – 1519)	between 1503 and 1507	(0.038,0.460)
Leonardo (1452 – 1519)	between 1503 and 1507	(0.029,0.564)
Leonardo da Vinci (1452 – 1519)	early 1500s	(0.019,0.562)
Leonardo (1452 – 1519)	early 1500s	(0.014,0.607)

5 Conclusions

In this paper we have shown how the formalism of prioritized fuzzy constraints allows to unify and generalize three approaches to estimate the feasibility of an answer candidate: frequency counts (e.g. Eq. (8)), co-occurrence statistics (e.g. Eq. (9)) and asking additional questions (e.g. Eq. (5)). The usefulness of representing answer candidates by fuzzy sets was illustrated by considering the problem of searching the creation date of a work of art, which in practice is often stated by means of an interval or a fuzzy description instead of an exact date.

Acknowledgements

Steven Schockaert and Martine De Cock would like to thank the Fund for Scientific Research – Flanders for funding their research.

References

1. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets and Systems **20** (1986) 87–96
2. Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A.: Data-intensive question answering. Proc. of the 10th TREC Conf., "http://trec.nist.gov/pubs.html" (2001)
3. Clarke, C.L.A., Cormack, G.V., Lynam, T.R.: Exploiting redundancy in question answering. Proc. of the 24th annual int. ACM SIGIR conf. on Research and Development in Information Retrieval (2001) 358 – 365
4. Dubois, D., Fargier, H., Prade, H.: The calculus of fuzzy restrictions as a basis for flexible constraint satisfaction. Proc. of the 2nd IEEE Int. Conf. on Fuzzy Systems (1993) 1131-1136
5. Harabagiu, S., Moldovan, D., Paşca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, R., Rus, V., Morărescu, P.: Falcon: boosting knowledge for answer engines. Proc. of 9th TREC Conf., "http://trec.nist.gov/pubs.html" (2000)
6. Kwok, C.C.T., Etzioni, O., Weld, D.S.: Scaling question answering to the web. Proc. of the 10th WWW Conf. (2001) 150–161
7. Luo, X., Lee, J.H-m., Leung, H-f., Jennings, N.R.: Prioritised fuzzy constraint satisfaction problems: axioms, instantiation and validation. Fuzzy Sets and Systems **136** (2003) 151–188

8. Novák, V., Perfilieva, I., Močkoř, J.: Mathematical Principles of Fuzzy Logic. Kluwer Academic Publishers (1999)
9. Prager, J., Chu-Carroll, J., Czuba, K.: Question answering using constraint satisfaction: QA-by-Dossier-with-Constraints. Proc. of the 42nd Annual Meeting of the ACL (2004) 574 – 581
10. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning I. Information Sciences **8** (1975) 199–249