Answer Generation with Temporal Data Integration

Véronique Moriceau
Université Paul Sabatier - IRIT
31062 Toulouse cedex 09, France
moriceau@irit.fr

Abstract

In this paper, we propose an approach for content determination and surface generation of answers in a question-answering system on the web. The content determination is based on a *coherence rate* which takes into account coherence with other potential answers. Answer generation is made through the use of classical techniques and templates and is based on a *certainty degree*.

1 Introduction

Search engines on the web and most of existing question-answering systems provide the user with either a set of hyperlinks or web page extracts containing answer(s) to a question.

As *provenance* information (defined in [McGuinness et al., 2004] e.g., source, date, author, etc.) is rather difficult to obtain, we assume that all web pages are equally reliable. Then, the problem the system has to solve is to generate an answer to a question even if several possible answers are selected by the extraction engine. For this purpose, we propose to integrate, according to certain criteria, the different possible answers in order to generate a single coherent answer which take into account the diversity of answers (which can be redundant, incomplete, inconsistent, etc.).

As our framework is WEBCOOP [Benamara, 2004], a co-operative question-answering system on the web, our goal is to generate answers in natural language which explain how confident of the answer the user can be.

In this paper, we focus on aspects of content determination and on the generation of answers in natural language. In the following sections, we first present the main difficulties and a general typology of integration mechanisms. Then we analyse the content determination process in the case of answers of type *date*. Finally, we present briefly a few elements about generation of integrated answers and evaluation.

2 Motivations

When a user submits a question to a classical search engine or question-answering system, he may obtain a set of potential answers which may be incoherent to some degree: we mean by incoherent, answers that are a priori contradictory but which can be in fact equivalent, complementary, etc. In this case, the user may be unsastisfied because he does not know which answer among those proposed is the correct one.

In the following sections, we present related works and a general typology of relations between candidate answers.

2.1 Related works

Most of existing systems on the web produce a set of answers to a question in the form of hyperlinks or page extracts, ranked according to a relevance score (for example, COGEX [Moldovan et al., 2003]). Other systems also define relationships between web page extracts or texts containing possible answers ([Harabagiu et al., 2004], [Radev et al., 1998]). For example, [Webber et al., 2002] defines 4 relationships between possible answers:

- **equivalence**: equivalent answers which entail mutually,
- **inclusion**: one-way entailment of answers,
- **aggregation**: answers that are mutually consistent but not entailing, and that can be replaced by their conjunction,
- **alternative**: answers that are inconsistent or alternatives and that can be replaced by their disjunction.

Most of question-answering systems generate answers which take into account neither information given by all candidate answers nor their inconsistency. This is the point we focus on in the following section.

2.2 A general typology of integration mechanisms

To better characterise our problem, we collected, via Google or QRISTAL [QRISTAL], a corpus of around 100 question-answer pairs in French that reflect different inconsistency problems. We first assume that all candidate answers are potentially correct. The corpus analysis enables us to define a general typology of relations between answers. For each relation defined in [Webber et al., 2002], we identify **integration** mechanisms in order to generate answers which take into account characteristics of all candidate answers.

Inclusion

The inclusion relation exists if a candidate answer entails another answer (for example, between concepts of candidate answers linked in an ontology by the *is-a* or *part-of* relations).

For example, *in Brittany* and *in France* are correct answers to the question *Where is Brest?* and *Brittany* is a part of *France*. The content determination stage consists here in choosing which answer will be proposed to the user - the more specific, the more generic or all answers. This can be guided by a user model, taking into account his knowledge.

Equivalence

Candidate answers which are linked by an equivalence relation are consistent and entail mutually. The corpus analysis allows us to identify two main types of equivalence:

(1) Lexical equivalence: synonymy, metonymy, paraphrases, proportional series, use of acronyms or foreign languages. For example, to the question *Who killed John Lennon?, Mark Chapman, the murderer of John Lennon* and *John Lennon's killer Mark Chapman* are equivalent answers.

(2) Equivalence with inference: in a number of cases, some common knowledge, inferences or calculation are necessary to detect equivalence relations. For example, *The A320 is 21* and *The A320 has been created in 1984* are equivalent answers to the question *How old is the Airbus A320?*.

Aggregation

The aggregation relation defines a set of consistent answers when the question accepts several different ones. In this case, all candidate answers are potentially correct and can be integrated in the form of a conjunction of all these answers. For example, an answer to the question *Where is Disneyland?* can be *in Tokyo, Paris, Hong-Kong and Los Angeles*.

If answers are numerical values, the integrated answer can be given in the form of an interval, average or comparison.

Alternative

The alternative relation defines a set of inconsistent answers. In the case of questions expecting a unique answer, only one answer among candidates is correct. On the contrary, all candidates can be correct answers.

(1) A simple solution is to propose a disjunction of candidate answers. For example, if the question *When does autumn begin?* has the candidate answers *Autumn begin on September 21st* and *Autumn begins on September 20th*, an answer such as *Autumn begins on either September 20th or September 21st* can be proposed.

(2) If candidate answers have common characteristics, it is possible to integrate them according to these characteristics. For example, the question *When does the French music festival take place?* has the following answers *June 1st 1982, June 21st 1983, ..., June 21st 2004*. Here, the extraction engine selects pages containing the dates of all music festivals. These candidate answers have day and month in common. Consequently, an answer such as *The French music festival takes place every June 21st* can be proposed.

(3) As for the aggregation relation, numerical values can be integrated in the form of an interval, average or comparison. For example, if the question *How far is Paris from Toulouse?* has the candidate answers *713 km, 678 km* and *681 km*, answers such as *Paris is at about 690 km from Toulouse* (average) or *The distance between Paris and Toulouse is between 678 and 713 km* (interval) can be proposed.

In the following sections, we focus on the content determination and generation of candidate answers of type *date* linked by an aggregation or alternative relation, the most common ones.

3 Content determination

The problem we focus on in this section is the problem of content determination when several answers to a question of type *date* are selected. We consider that candidate answers can be in the form of *date* or temporal *interval*. A *date* is defined as a vector which allows the temporal localisation of an event. Some values of vectors can be underspecified: only relevant values for the expected information are explicit (year, hour, etc.). Then, an *interval* is a couple of *dates*, i.e. vectors defining a date of beginning and a date of end.

As answers selected by the extraction engine are often in different forms (dates or intervals or both), a first step consists in standardizing data:

- all candidate answers are in the form of an *interval*: this means that a *date* will be in the form of an *interval* having the same date of beginning and of end,
- some candidate answers may be incomplete: for example, year or date of end is missing, etc. In some cases, unification with other candidate answers is possible. Otherwise, incomplete answers are omitted,
- from the semantic point of view, all candidate answers must be in the same system of temporal reference (for example, because of possible different time zones).

Once all candidate answers have been standardized, aberrant answers are filtered out by applying classical statistical methods. Then, the answer selection process can be applied.

3.1 Answer selection process

Our goal is to select, among several candidate answers, the *best* answer considered as the one which is the most coherent with other answers. For this purpose, we define a **coherence rate** of answers.

Let us assume that there are N candidate answers coming from N different web pages. We consider that each candidate answer is a temporal interval $[d_b, d_e]$ where d_b is the date of beginning and d_e the date of end of the event. Let $d_i = [d_{b_i}, d_{e_i}]$ with $1 \leq i \leq N$ be these N candidate answers.

In terms of interval, we consider that the most coherent answer is the interval which intersects the greatest number of candidate intervals. For example, in Figure 1, we have 3 candidate answers d_1, d_2 and d_3. They form 4 sub-intervals: $[d_{b_1}, d_{b_2}]$, $[d_{b_2}, d_{b_3}]$, $[d_{b_3}, d_{e_3}]$ and $[d_{e_3}, d_{e_1}]$.

The interval we consider as the most coherent is $[d_{b_3}, d_{e_3}]$ because its occurrence frequency is 3 (i.e. the number of times it intersects the candidate answers is 3).

In order to define sub-intervals, we need to have the bounds of the N candidate intervals. Let $B = \{d_{b_j}, d_{e_j}\}$, 1$\leq j \leq$N, be the set of ordered bounds of the N intervals and let $m_i \in B$, 1$\leq i \leq$2N. Consequently, a sub-interval is in the form of $[m_i, m_{i+1}]$.

We now define F_k as the occurrence frequency of the

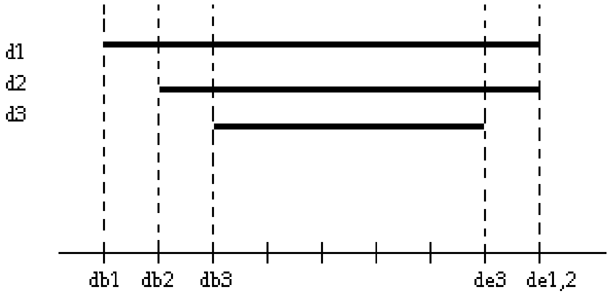


Figure 1: Sub-intervals

interval d_k, i.e. the number of times d_k intersects the N candidate answers:
$\forall\ 1 \leq i \leq 2N-1,\ F_{[m_i,m_{i+1}]} =\ card\ (\ \{\ [m_i, m_{i+1}],$
$such\ as\ \forall\ 1 \leq j \leq N,\ \ [m_i, m_{i+1}] \subseteq d_j\ \})$

Then, the *coherence rate* t_i assigned to each sub-interval $[m_i, m_{i+1}]$ is a weighting of the occurrence frequency by the number of candidate answers:

$$\forall\ 1 \leq i \leq 2N-1,\ \ t_i = \frac{F_{[m_i,m_{i+1}]}}{N}$$

Selecting the interval having the highest *coherence rate* is not sufficient. The answer must also have a relevant duration. For this purpose, we construct new intervals based on previous sub-intervals: these new ones must have a relevant duration, at least equal to the *average duration* of the N candidate answers. Let du_{av} be the *average duration* of candidate answers.

Then, we construct a coherent answer set composed of intervals satisfying a constraint duration to which we assigned a new *coherence rate*. This new rate is the average of the coherence rates of sub-intervals composing the new one. So, the coherent answer set A is defined as:

$$A = \{\ ([m_i, m_{k+1}], t_{i,k+1}),\ \ \forall\ 1 \leq i \leq 2N-1,$$

$$i < k \leq 2N-1, \quad [m_i, m_{k+1}] = \bigcup_{j=i}^{k} [m_j, m_{j+1}]$$

$$with\ \ t_{i,k+1} = \sum_{j=i}^{k} \frac{t_j}{k+1-i},\ \ such\ as$$

$$Int(du_{av})\ \leq du_{[m_i,m_{k+1}]}\ \leq\ Int(du_{av}) + 1\ \ \}$$

Once this coherent answer set has been obtained, there is still to check if the expected answer/event is a unique or an iterative event. We consider that an event is iterative if there is a great number of intervals of A that are distant in time. Let α be the minimum time between the end of an interval and the beginning of the following one. Let β be the minimum number of intervals that have to be α distant from the others (the parameters α and β depends on data granularity). Then, an event is iterative if:

$\forall\ 1 \leq i \leq 2N-1,\ \ i < k \leq 2N-1,$
$card\ (\ \{\ [m_i, m_k] \in A\ \text{ such as }\ \forall\ i{+}1 \leq j \leq 2N-1,$
$[m_{i+1}, m_j] \in A,\ \ |m_{i+1} - m_k| \geq \alpha\ \}\ )\ \geq\ \beta$

At this stage, there are two possibilities:

- either the event is unique: the answer set Ans is composed of intervals of A having the highest *coherence rate*:
  $Ans = \{\ ([m_i, m_k], t_{ik}),\ \ 1 \leq i \leq 2N\text{-}1,\ such\ as$
  $\forall\ ([m_j, m_l], t_{jl}) \in A,\ 1 \leq j \leq 2\text{N-}1,\ t_{ik} = max(t_{jl})\ \}$
- or the event is iterative: there may be some temporal constraints due to the question: for example, the question expects an event in the past or in the future, an event in a particular year, etc. Let A_q be the set of intervals of A satisfying the question constraints. Then, Ans is the set of answers/intervals (having the highest coherence rate) which can be proposed to the user:
  $Ans = \{\ ([m_i, m_k], t_{ik}),\ \ 1 \leq i \leq 2N\text{-}1,\ such\ as$
  $\forall\ ([m_j, m_l], t_{jl}) \in A_q,\ 1 \leq j \leq 2\text{N-}1,\ t_{ik} = max(t_{jl})\ \}$

In this section, we proposed a method for content determination based on *coherence rate* in the case of answers of type *date* and in particular of type *interval*. In the following section, we apply this method to an example.

3.2 Example

Let us suppose that the question *When did Hugo hurricane take place?* is submitted to a question-answering system. The following table presents the candidate answers:

Question	*When did Hugo hurricane take place?*
Candidate Answers	September 16th, 1989
	September 1989, from 10 to 22
	September 16th, 1989
	September 17th, 1989
	from 10th to 25th September, 1989
	September 16th, 1989
	September 16th, 1989
	from 16th to 22nd September, 1989
	September 1989, from 10 to 25
	September 16th, 1989
	September 16th, 1989

The following table presents the 11 candidate answers in the form of interval and their respective duration (number of days):

Question	*When did Hugo hurricane take place?*
Candidate Answers	d_1 = [16-9-1989,16-9-1989], du_1 = 1
	d_2 = [10-9-1989,22-9-1989], du_2 = 12
	d_3 = [16-9-1989,16-9-1989], du_3 = 1
	d_4 = [17-9-1989,17-9-1989], du_4 = 1
	d_5 = [10-9-1989,25-9-1989], du_5 = 15
	d_6 = [16-9-1989,16-9-1989], du_6 = 1
	d_7 = [16-9-1989,16-9-1989], du_7 = 1
	d_8 = [16-9-1989,22-9-1989], du_8 = 6
	d_9 = [10-9-1989,25-9-1989], du_9 = 15
	d_{10} = [16-9-1989,16-9-1989], du_{10} = 1
	d_{11} = [16-9-1989,16-9-1989], du_{11} = 1

The ordered set of interval bounds is for example: $B = \{ \, d_{b_2}, \, d_{b_1}, \, d_{e_1}, \, d_{b_4}, \, d_{e_4}, \, d_{e_3}, \, d_{e_5} \, \}$

Consequently, we have (cf. Figure 2):
$m_1 = d_{b_2}$ = 10-9-1989, $m_2 = d_{b_1}$ = 16-9-1989,
$m_3 = d_{e_1}$ = 16-9-1989, $m_4 = d_{b_4}$ = 17-9-1989,
$m_5 = d_{e_4}$ = 17-9-1989, $m_6 = d_{e_3}$ = 22-9-1989,
$m_7 = d_{e_5}$ = 25-9-1989,

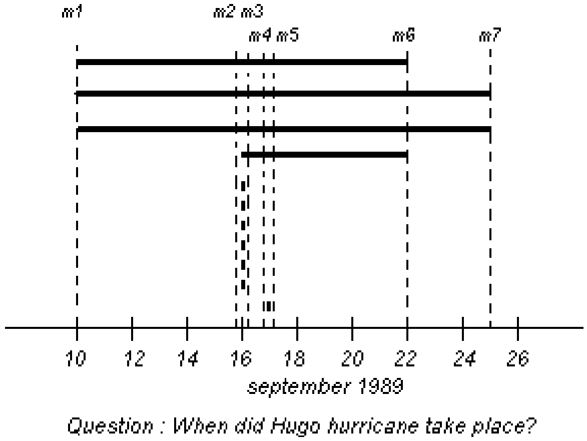


Figure 2: 11 candidate answers

The *coherence rates* of each sub-interval are:

$$t_1 = \frac{F_{[m_1,m_2[}}{N} = \frac{card([m_1,m_2[\subset d_{j_{(1\leq j\leq N)}})}{N} = \frac{3}{11} = 0.27$$

$$t_2 = \frac{F_{[m_2,m_3]}}{N} = \frac{card([m_2,m_3] \subset d_{j_{(1\leq j\leq N)}})}{N} = \frac{10}{11} = 0.91$$

$$t_3 = \frac{F_{]m_3,m_4[}}{N} = \frac{card(]m_3,m_4[\subset d_{j_{(1\leq j\leq N)}})}{N} = \frac{4}{11} = 0.36$$

$$t_4 = \frac{F_{[m_4,m_5]}}{N} = \frac{card([m_4,m_5] \subset d_{j_{(1\leq j\leq N)}})}{N} = \frac{5}{11} = 0.45$$

$$t_5 = \frac{F_{]m_5,m_6[}}{N} = \frac{card(]m_5,m_6[\subset d_{j_{(1\leq j\leq N)}})}{N} = \frac{4}{11} = 0.36$$

$$t_6 = \frac{F_{[m_6,m_7]}}{N} = \frac{card([m_6,m_7] \subset d_{j_{(1\leq j\leq N)}})}{N} = \frac{2}{11} = 0.18$$

The average duration of candidate answers is 5 days. Now, we construct the answer set A with sub-intervals having a duration between 5 and 6 days and we assign to them a new *coherence rate*:

$$du_{[m_1,m_2[} = 5 \;\; and \;\; t_{12} = t_1 = 0.27$$
$$du_{[m_1,m_3]} = 6 \;\; and \;\; t_{13} = \frac{t_1 + t_2}{2} = 0.59$$
$$du_{[m_2,m_6[} = 6 \;\; and \;\; t_{26} = \frac{t_2 + t_3 + t_4 + t_5}{4} = 0.52$$
$$du_{]m_3,m_6[} = 6 \;\; and \;\; t_{36} = \frac{t_3 + t_4 + t_5}{3} = 0.39$$
$$du_{[m_4,m_6[} = 5 \;\; and \;\; t_{46} = \frac{t_4 + t_5}{2} = 0.41$$
$$du_{]m_5,m_6[} = 5 \;\; and \;\; t_{56} = t_5 = 0.36$$

Consequently, the intervals satisfying the average duration are: $A = \{ \, [m_1, m_2[, \; [m_1, m_3], \; [m_2, m_6[, \;]m_3, m_6[, \; [m_4, m_6[, \;]m_5, m_6[\, \}$

The event is non-iterative since every interval of A is contiguous to the following one. So, the answer is the interval of A having the highest *coherence rate*: $Ans = ([m_1, m_3], \; 0.59)$ i.e. from September, 10th to 16nd 1989.

4 Answer generation

Once the most coherent answer has been elaborated, it has to be generated in natural language. Our strategy is to couple classical NLG techniques with generation templates.
As our framework is the cooperative system WEBCOOP, the answer proposed to the user has to explain why this answer has been selected. The idea is to introduce possibility degrees to explain to the user how confident of the answer he can be. For this purpose, we define a **certainty degree** of answers which depends on several parameters:

- the number of candidate answers (N): if N and the *coherence rate* of the selected answer are high, then this means that there were not many contradictions among candidate answers and that the answer is more certain (as N is already taken into account in the *coherence rate*, only this rate is a sufficient parameter),
- if the difference τ between the best coherence rate and the second best one is high, then this means that the selected answer is more certain.

Consequently, we define the *certainty degree* δ_{ik} of the answer $[m_i, m_k]$ as:

$$\delta_{ik} = \begin{cases} 1 & \text{if} \quad t_{ik} = 1 \\ \tau \times t_{ik} & \text{with :} \end{cases}$$

$([m_i, m_k], t_{ik}) \in Ans$ and $\tau = t_{ik} - t_{jl}$ where t_{ik} is the best coherence rate and t_{jl} the second best one.

As $0 \leq t_{ik} \leq 1$ and $0 \leq \tau \leq 1$, the more δ_{ik} tends towards 1, the more the answer $[m_i, m_k]$ is certain. Thus, we define generation schemas for each type of answer depending on this *certainty degree*. We distinguish 3 main cases:
(1) either $Ans = \emptyset$, i.e. no answer has been selected. The idea is to select the candidate answer which has the highest *coherence rate* even if its duration is not appropriate but the generated answer has to explain that this answer is not sure,

(2) or $\delta_{ik} = 1$, i.e. the selected answer $[m_i, m_k]$ is certain,
(3) or $\delta_{ik} \neq 1$, then the generated answer has to take into account τ. If τ is low, the *coherence rate* of the selected answer is very close to other rates: in this case, several answers are potentially correct and can be proposed to the user.

The idea is to generate answers with different certainty degrees depending on δ: we choose to express this degree by the use of adverbs. For this purpose, we define a lexicalisation function *lex* which lexicalises the selected answers and a function *lexD* which lexicalises δ. The Table 1 presents the different generation schemas (A is the selected answer and A' the answer having the *coherence rate* the closest to δ_A). Underlined fragments are predefined texts.

case (1)	subject lexD(δ_A, min) verb lex(A, Reg)
case (2)	subject verb lex(A, Reg)
case (3)	**τ is high:** subject lexD(δ_A, _) verb lex(A, Reg) **τ is low:** A and A' are proposed *if A is a date:* subject lexD(δ_A, _) verb lex(A, Reg) or lex(A′, Reg) *if A is an interval:* subject lexD($\delta_{A'}$, _) verb lex(A′, Reg) but lexD($\delta_{A'}$, plus) lex(A, Reg)

Table 1: Generation schemas

Adverb intensity is represented by the following proportional serie (cf. Figure 3):

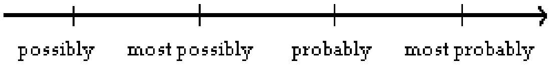


Figure 3: Adverb intensity

Consequently, if δ is high, it will be lexicalised by an adverb of high intensity. The second argument of the function $lexD$ ($minus$ or $plus$) forces the function to lexicalise δ as an adverb of lower or higher intensity than the one that would have been used normally (case (1) and (3)).

The lex function has 2 arguments: the answers that have to be generated and Reg indicating if the event is regular or not. Indeed, if an iterative event is regular, i.e. happens at regular intervals (i.e. the parameter α is always the same for all answers of A), then generalisation can be made on common characteristics. For example, if α = 1 year, a possible generalisation is: *X takes place* **every year** *on* ....

Example 1

To the question *When was Chomsky born?*, the only potential answer and its respective coherence rate is ([07-12-1928, 07-12-1928], 1). Its certainty degree is: $\delta = 1$.
We are in case (2) so the generated answer is in the form:
subject verb lex(A, Reg).
The answer is not a regular event. Consequently, the answer in natural language is:
Chomsky was born **on December, 7th 1928**.

Example 2

To the question *In which year did D. Tutu receive the Nobel Peace Prize?*, the potential answers and their respective coherence rate are: (1931, 0.08), (1984, 0.87) and (1986, 0.04). The answer (1984, 0.87) is selected because it has the highest coherence rate and its certainty degree is:
$\delta = (0.87 - 0.08) \times 0.87 = 0.69$
We are in case (3) with a high τ $(0.87 - 0.08)$ so the generated answer is in the form:
subject lexD(δ_A, _) verb lex(A, Reg).
The answer is not a regular event and its certainty degree is high so the adverb intensity has to be high. Consequently, the answer in natural language is:
D. Tutu **probably** *received the Nobel Peace Prize* **in 1984**.

Example 3

To the question *When did the American Civil War take place?*, the potential answers and their respective coherence rate are:
- ([01-01-1861, 09-04-1865], 0.29),
- ([12-04-1861, 09-04-1865], 0.32),
- ([17-04-1861, 09-04-1865], 0.33).

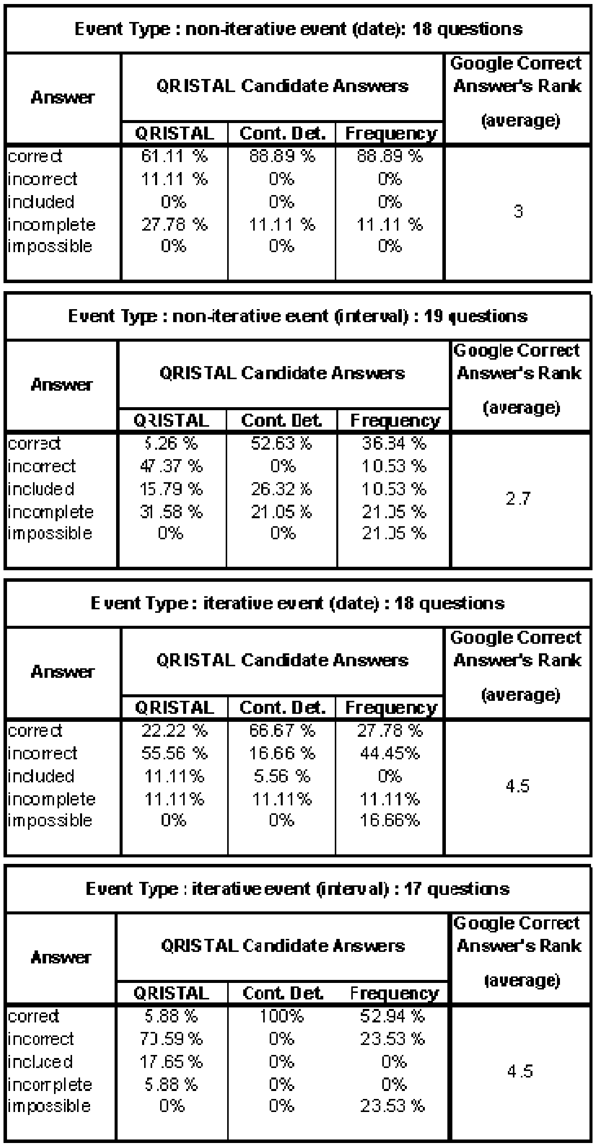
The answer ([17-04-1861, 09-04-1865], 0.33) is selected because it has the highest coherence rate and its certainty degree is: $\delta = (0.33 - 0.32) \times 0.33 = 0.003$
We are in case (3) with a low τ $(0.33 - 0.32)$ and the answer is an interval so the generated answer is in the form:
subject lexD($\delta_{A'}$, _) verb lex(A′, Reg) but lexD($\delta_{A'}$, plus) lex(A, Reg),
with A' = [01-01-1861, 09-04-1865] (since all other answers have a quasi-similar coherence rate, A' is the interval including all the others). The answer is not a regular event and its certainty degree is very low so the adverb intensity has to be very low. Consequently, the answer in natural language is:
The American Civil War **possibly** *took place* **from 1861 to April, 9th 1865** *but* **most possibly from April, 17th 1861 to April, 9th 1865**.

In this paper, we did not detail the lexicalisation of dates but classical lexicalisation and aggregation techniques are applied for example to group common characteristics (*from September, 10th to 22th* instead of *from September, 10th to September, 22th*, etc).

5 Evaluation

We evaluate our approach by applying our answer selection method to 72 questions expecting an answer of type *date*. Among these questions, 36 questions expected an answer of type *date* and 36 expected an *temporal interval*.
These 72 questions were submitted to QRISTAL. Applying

our answer selection process (called *Cont.Det.* in the following tables), we distinguish several cases: either the proposed answer is correct, or it is incorrect or the proposed answer is included in the interval defining the exact date of the event or the answer is incomplete. We note ”impossible” cases when it is impossible to select an answer (when all candidate answers have the same occurrence frequency).



Event Type : non-iterative event (date): 18 questions				
Answer	QRISTAL Candidate Answers			Google Correct Answer's Rank (average)
	QRISTAL	Cont. Det.	Frequency	
correct	61.11 %	88.89 %	88.89 %	3
incorrect	11.11 %	0%	0%	
included	0%	0%	0%	
incomplete	27.78 %	11.11 %	11.11 %	
impossible	0%	0%	0%	

Event Type : non-iterative event (interval) : 19 questions				
Answer	QRISTAL Candidate Answers			Google Correct Answer's Rank (average)
	QRISTAL	Cont. Det.	Frequency	
correct	5.26 %	52.63 %	36.34 %	2.7
incorrect	47.37 %	0%	10.53 %	
included	15.79 %	26.32 %	10.53 %	
incomplete	31.58 %	21.05 %	21.05 %	
impossible	0%	0%	21.05 %	

Event Type : iterative event (date) : 18 questions				
Answer	QRISTAL Candidate Answers			Google Correct Answer's Rank (average)
	QRISTAL	Cont. Det.	Frequency	
correct	22.22 %	66.67 %	27.78 %	4.5
incorrect	55.56 %	16.66 %	44.45%	
included	11.11%	5.56 %	0%	
incomplete	11.11%	11.11%	11.11%	
impossible	0%	0%	16.66%	

Event Type : iterative event (interval) : 17 questions				
Answer	QRISTAL Candidate Answers			Google Correct Answer's Rank (average)
	QRISTAL	Cont. Det.	Frequency	
correct	5.88 %	100%	52.94 %	4.5
incorrect	70.59 %	0%	23.53 %	
included	17.65 %	0%	0%	
incomplete	5.88 %	0%	0%	
impossible	0%	0%	23.53 %	

Figure 4: Evaluation on 72 questions

We compare the results of our content determination method not only to QRISTAL's results but also to the results obtained by a ”*most frequent answer*” method. Our approach obtains better results on questions expecting an answer of type *temporal interval* and particularly on questions about iterative events (for example, *When does the next X take place? When did the first Y happen?*, ...). This is partly due to the fact that a ”*most frequent answer*” method, for example, is not able to solve temporal references.

Among the ”incorrect” answers, most errors can be explained by the fact that some incorrect candidate answers introduce a bias in the calculation of the average duration. A way to solve this problem is to eliminate some candidate answers by analysing in more depth their contexts of occurrence. Linguistic information and semantic knowledge about answer concepts may allow to determine if a candidate answer selected by QRISTAL is appropriate or not, incomplete, etc.

6 Conclusion

In this paper, we presented an approach for content determination, based on a *coherence rate*, and surface generation, based on a *certainty degree* of answers in a question-answering system on the web. Several future directions are obviously considered:

- analyse in more depth of the contexts of occurrence of candidate answers in order to filter out incorrect answers or to precise some of them. This analysis will avoid having answers which introduce a bias in calculations,
- evaluation of the quality of answers in natural language: are adverbs sufficient to explain the certainty degree of the answer?.

References

[Benamara, 2004] F. Benamara. *WEBCOOP: un système question-réponse coopératif sur le Web.* PhD Thesis, Université Paul Sabatier, Toulouse, 2004.

[Chalendar et al., 2002] G. de Chalendar, T. Delmas, F. Elkateb, O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, A. Vilnat. *The Question-Answering system QALC at LIMSI, Experiments in using Web and WordNet.* In Proceedings of TREC 11, 2002.

[Harabagiu et al., 2004] S. Harabagiu, F. Lacatusu. *Strategies for Advanced Question Answering.* In Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004.

[McGuinness et al., 2004] D.L. McGuinness, P. Pinheiro da Silva. *Trusting Answers on the Web.* New Directions in Question-Answering, chapter 22, Mark T. Maybury (ed), AAAI/MIT Press, 2004.

[Moldovan et al., 2003] D. Moldovan, C. Clark, S. Harabagiu, S. Maiorano. *COGEX: A Logic Prover for Question Answering.* In Proceedings of HLT-NAACL 2003.

[Radev et al., 1998] D.R. Radev, K.R. McKeown. *Generating Natural Language Summaries from Multiple On-Line Sources.* Computational Linguistics, vol. 24, issue 3 - Natural Language Generation, pp. 469 - 500, 1998.

[Webber et al. 2002] B. Webber, C. Gardent, J. Bos. *Position statement: Inference in Question Answering.* In Proceedings of LREC, 2002.

[QRISTAL] Question-Réponse Intégrant un Système de Traitement Automatique des Langues. www.qristal.fr, Synapse Développement, 2004.