Language Challenges for Data Fusion in Question-Answering

Véronique Moriceau

Institut de Recherche en Informatique de Toulouse

118, route de Narbonne

31062 Toulouse cedex 9, France

E-mail: moriceau@irit.fr

Abstract

Search engines on the web and most existing question-answering systems provide the user with a set of hyperlinks and/or web page extracts containing answer(s) to a question. These answers are often incoherent to a certain degree (equivalent, contradictory, etc.). It is then quite difficult for the user to know which answer is the correct one. In this paper, we present an approach which aims at providing synthetic numerical answers in a question-answering system. These answers are generated in natural language and, in a cooperative perspective, the aim is to explain to the user the variation of numerical values when several values, apparently incoherent, are extracted from the web as possible answers to a question. We present in particular how lexical resources are essential to answer extraction from the web, to the characterization of the variation mode associated with the type of information and to answer generation in natural language.

1. Introduction

Search engines on the web and most existing question-answering systems provide the user with a set of hyperlinks and/or web page extracts containing answer(s) to a question. These answers may be incoherent to a certain degree: they may be equivalent, complementary, contradictory, at different levels of precision or specificity, etc. The user has then to select and read pages to find an answer: it is a quite long procedure and it is also difficult for the user to know which answer is the correct one.

Some systems define relationships between web page extracts or texts containing possible answers: for example, (Radev and McKeown, 1998) and (Harabagiu and Lacatusu, 2004) define *agreement* (when two sources report the same information), *addition* (when a second source reports additional information), *contradiction* (when two sources report conflicting information), etc. These relations can be classified into the 4 relations defined by (Webber et al., 2002), i.e. **inclusion** (a candidate answer is in an inclusion relation if it entails another answer), **equivalence** (candidate answers which are linked by an equivalence relation are consistent and entail mutually), **aggregation** (it defines a set of consistent answers when the question accepts several different ones) and **alternative** (it defines a set of inconsistent answers).
Most question-answering systems provide answers which take into account neither information given by all candidate answers nor their inconsistency. This is the point we focus on.

In a cooperative perspective as defined in (Grice, 1975) (*be as informative as necessary, do not make your contribution to the conversation more informative than necessary*, …), we propose an approach for answer generation in natural language which uses answer **integration**. When several possible answers are selected by the extraction engine, the goal is to define a coherent core from candidate answers and to generate a **cooperative answer**, i.e. an answer with explanations. We assume that all web pages are equally reliable since page *provenance* information (defined in (McGuinness and Pinheiro da Silva, 2004) e.g., source, date, author, etc.) is difficult to obtain.

In this paper, we focus on questions expecting answers of type n*umerical* and explain how lexical resources are essential to:

- answer extraction from the web,
- the discovery of the variation mode associated with the type of information and
- answer generation in natural language (in French[1]).

2. Motivations

Numerical questions deal with numerical properties such as distance, quantity, weight, age, etc. In order to identify the different problems, let us consider the following example.

How many inhabitants are there in France?
- 01/2000: France has officially 60186184 inhabitants.
- 61.7 millions of inhabitants in France in 2004.
- January, 19th 2005: 62 millions of inhabitants in France.

This set of potential answers may seem incoherent but their internal coherence can be made apparent once a variation criterion is identified (in this example, the number of inhabitants changes over time).
In a cooperative perspective, an answer can be for example:

In 2005, there are 62 millions of inhabitants in France. It increased by about 2 millions between 2000 and

[1] Examples are English glosses.

2005.
This answer is composed of:

1. a direct answer to the question,
2. an explanation characterizing the variation mode of the numerical value.

To generate this kind of answer, it is necessary (1) to integrate candidate answers in order to elaborate a direct answer (for example by solving inconsistencies), and (2) to integrate candidate answer characteristics in order to generate an explanation.

In the following sections, we first define a typology of numerical answers and then briefly present the general architecture of the system which generates cooperative numerical answers.

2.1 A Typology of Numerical Answers

To define the different types of numerical answers, we collected a set of 80 question-answer pairs about prices, quantities, age, time, weight, temperature, speed and distance. The goal is to identify for each question-answer pair why extracted numerical values are different (is this an inconsistency? an evolution?).
A question may have several correct numerical answers when numerical values vary according to certain criteria. Let us consider the following examples.

Example 1:
How many inhabitants are there in France?
- Population census in France (1999): 61632485.
- 61.7: number of inhabitants in France in 2004.
In this example, the numerical value (quantity) is a property which changes over time (*1999*, *2004*).

Example 2:
What is the average age of marriage of women in 2004?
- In Iran, the average age of marriage of women went from 19 to 21 years in 2004.
- In 2004, Moroccan women get married at the age of 27.
In this example, the numerical value (age of marriage) varies according to place (*in Iran*, *Moroccan*).

Example 3:
At what temperature should I serve wine?
- *Red wine must be served at room temperature.*
- *Champagne: between 8 and 10°C.*
- *White wine: between 8 and 11°C.*

Here, the numerical value (temperature) varies according to the question focus (type of *wine*).

The corpus analysis allows us to identify 3 main variation criteria, namely *time*, *place* and *restriction* (restriction on the focus, for example: Champagne/wine). These criteria can be combined: some numerical values vary according to time and place, to time and restrictions, etc. (for example, the average age of marriage vary according to time, place and restrictions on men/women).

2.2 Architecture of the System

Figure 1 presents the general architecture of our system which allows us to generate answers and explanations from several different numerical answers.

We use QRISTAL, a question-answering system on the web, as a question analyzer and a search engine. Questions are submitted in natural language and are analyzed syntactically (identification of keywords) and semantically (disambiguation, focus, answer expected type). Then QRISTAL selects potential answers from the web: it searches web pages containing the keywords of the query and synonyms (Laurent and Séguéla, 2005).
Then, an extraction grammar constructs a set of frames from candidate web pages. From the frame set, the variation criteria and mode of the searched numerical value are identified. Finally, a natural language answer is generated explaining those characteristics.
Each of these stages is presented in the next sections.

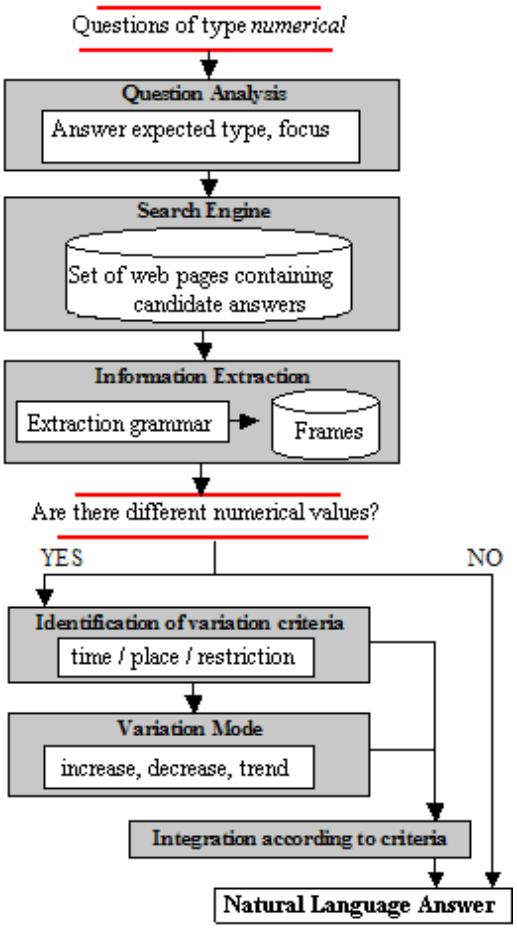

Figure 1 : Architecture of the System

3. Extraction and Characterization of Answers

Answer characterization consists in 2 main stages:
- information extraction from candidate web pages,
- characterization of variation (criteria and mode) of numerical values if necessary.

3.1 Answer Extraction

Once QRISTAL has selected candidate web pages, a grammar is applied to extract information needed for the generation of an appropriate cooperative answer.
This information is mainly:

- the searched numerical value (*val*),
- the *unit* of measure,
- the question focus and its synonyms (*focus*) ,

- the *date* and *place* of the information,
- the *restriction*(s) on the question focus (essentially, adjectives or relative clauses),

and linguistic clues indicating:

- the *precision* of the numerical value (for example adverbs or prepositions such as in *about 700*, ...),
- a *variation* of the value (for example temporal adverbs, verbs of change/movement as in *the price* ***increased*** *to 200 euro*).

All this information for a numerical value are gathered in a frame a_i. A dedicated grammar extracts this information from candidate web pages and produces the set *A* of *N* candidate answers: $A = \{a_1, ..., a_N\}$.

We use a gapping grammar (Dahl and Abramson, 1984) to skip elements which are not useful. We give below the main rules of the grammar, optional elements are between brackets:

```
Answer → Nominal Sentence | Verbal Sentence

Nominal Sentence → Focus (Restriction), ...,
(Date), ..., (Place), ..., (Precision) Val (Unit)

Verbal Sentence → Focus (Restriction), ...,
(Date), ..., (Place), ..., Verb, ..., (Precision)
Val (Unit)

Verb → VerbQuestion | Variation
VerbQuestion → count | estimate | weigh | ...
Variation → go up | decrease | ...
Precision → about | on average | ...
Place → Country | City | ...
Time → Date | Period | ...
Restriction → Adjective | Relative | ...
.......
```

Figure 2 shows an extraction result.

What is the price of a Peugeot 206?
(1) Ads (August 2005): sell Peugeot 206 diesel, 3400 km, 16100 €, City: Toulouse, Tel: ...
(2) December 2005:
Peugeot 206 (gas): 17200 €, 2700 km, Address: ... Paris

$a_1 =$ (Val = 16100; Precision = ∅; Unit = €; Focus = Peugeot 206; Date = August 2005; Place = Toulouse; Restriction = diesel, 3400 km; Variation = ∅)

$a_2 =$ (Val = 17200; Precision = ∅; Unit = €; Focus = Peugeot 206; Date = December 2005; Place = Paris; Restriction = gas, 2700 km; Variation = ∅)

Figure 2 : Extraction Results

For verb identification, we use a classification of French verbs[2] (Saint-Dizier, 1999) based on the main classes defined by WordNet. The classes we are interested in for our task are mainly those of verbs of change (*increase*, *decrease*, etc.: in total, 262 verbs in French) and of verbs of movement (*climb*, *move forward/backward*, etc.: in total, 252 verbs in French) used metaphorically (Moriceau and Saint-Dizier, 2003). From theses classes, we collected a set of 74 verbs which can be applied to numerical values.

In the same way, for the extraction of *precision* information, we use PrepNet[3] (Saint-Dizier, 2005) which provides a relatively deep description of preposition syntactic and semantic behaviours. In particular, we are interested in prepositions of the class of quantity (precise numerical quantity and approximate quantity: about 15 prepositions in French).
Our grammar rules are based on the grammar defined in (Maurel, 1991) for date extraction and on an ontology of geographical places (cf. figure 3) for place information extraction.

3.2 Variation Characterization

Variation Criteria

The goal is to determine if there is a numerical variation and to identify the variation criteria of the value. In fact, we assume that there is a variation if there is at least *k* different numerical values with different criteria (time, place, restriction) among the *N* frames. Thus, a numerical value varies according to: (for more details, see (Moriceau, 2006))

(1) **time** if card ({a_i, such as $\exists\ a_i, a_j \subset A$,
a_i (Val) $\neq a_j$ (Val)
$\wedge$ a_i (Unit) = a_j (Unit)
$\wedge$ a_i (Date) $\neq a_j$ (Date) }) $\geq k$

(2) **place** if card ({a_i , such as $\exists\ a_i, a_j \subset A$,
a_i (Val) $\neq a_j$ (Val)
$\wedge$ a_i (Unit) = a_j (Unit)
$\wedge$ a_i (Place) $\neq a_j$ (Place) }) $\geq k$

(3) **restriction** if card ({a_i , such as $\exists\ a_i, a_j \subset A$,
a_i (Val) $\neq a_j$ (Val)
$\wedge$ a_i (Unit) = a_j (Unit)
$\wedge$ a_i (Restriction) $\neq a_j$ (Restriction) }) $\geq k$

(4) **time and place** if (1) $\wedge$ (2)
(5) **time and restriction** if (1) $\wedge$ (3)
(6) **place and restriction** if (2) $\wedge$ (3)
(7) **time, place and restriction** if (1) $\wedge$ (2) $\wedge$ (3)

In the example of figure 2, the price varies according to time, place and restriction.

Numerical values can be compared only if they have the same unit of measure. If not, they have to be converted.

[2] http://www.irit.fr/recherches/ILPL/essais/verbe.php

[3] http://www.irit.fr/recherches/ILPL/prepnet.html

For each criterion (time, place or restriction), only information of the same semantic type and of the same ontological level can be compared. For example, *population of overseas regions* and *metropolitan population* are restrictions of the same ontological type/ level and can be compared. On the contrary, *metropolitan population* and *prison population* are restrictions of a different ontological level and cannot be compared. In the same way, place criteria can only be compared if they have the same ontological level: for example, prices *in Paris* and *in Toulouse* can be compared because the ontological level of both places is *city*. On the contrary, prices *in Paris* and *in France* cannot be compared since the ontological levels are respectively *city* and *country* (cf. figure 3).

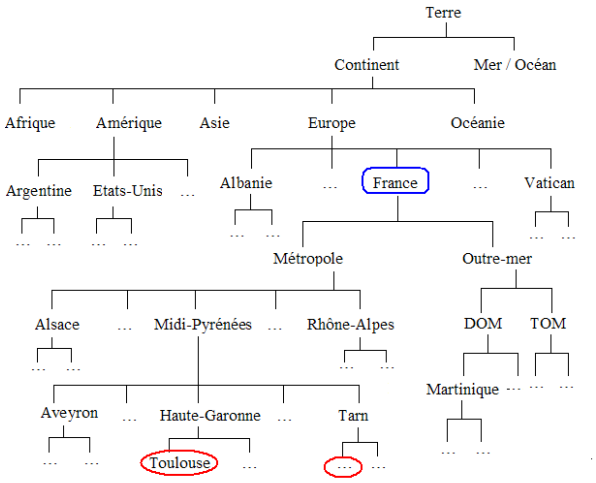

Figure 3 : Ontology of geographical places

In the following sections, we focus on numerical values which vary according to time.

Variation Mode

The last step consists in identifying the variation mode of values. The idea is to draw a trend (increase, decrease, ...) of variation in time so that an explanation can be generated: we draw a regression line which determines the relationship between the two extracted variables *numerical value* and *date*.

Pearson's correlation (*r*) reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A positive Pearson's correlation implies a general increase of values (trend) whereas a negative Pearson's correlation implies a general decrease. On the contrary, if *r* is low (-0.6 < r < 0.6), then the trend is mathematically considered as random (Fisher, 1925).

Figure 4 shows the results for the question *How many inhabitants are there in France?* The Pearson's correlation is 0.694 meaning that the number of inhabitants increases according to time (between 1999 and 2005).

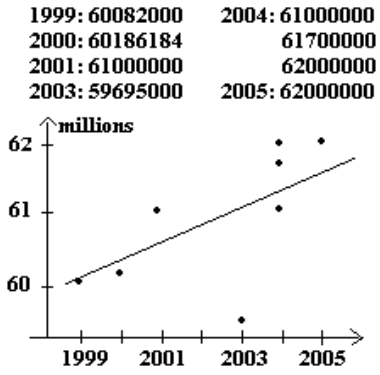

Figure 4 : Variation mode

4. Answer Generation

Once the searched numerical values have been extracted and characterized by their variation criteria and mode, a cooperative answer is generated in natural language. It is composed of two parts:

1. a direct answer if available,

2. an explanation of the value variation.

In the following sections, we present some prerequisites to the construction of each of these parts in term of resources and knowledge.

4.1 Direct Answer Generation

There are mainly two cases: either one or several criteria are constrained by the question (as in *How many inhabitants are there in France in 2005?* where criteria of place and time are given), or some criteria are omitted (or implicit, as in *How many inhabitants are there in France?* where there is no information on time). In the first case, the numerical value satisfying the constraints is chosen (unification between the criteria of the question and those extracted from web pages). In the second case, we assume that the user wants to have the most recent information.

We focus here on answers which vary according to time. Aberrant values are first filtered out by applying classical statistical methods. Then, when there is only one numerical value which satisfies the temporal constraint (given by the question or the most recent date), then the direct answer is generated from this value. When there is no numerical value satisfying the temporal constraint, only the second part of the answer (explanation) is generated.

In the case of several numerical values satisfying the temporal constraint, there may be approximate values. For example, the following answers (cf figure 4) are extracted for the question *How many inhabitants were there in France in 2004?*:

(1) 61.7 millions: number of inhabitants in France in 2004.

(2) In 2004, the French population is estimated to 61

millions.

(3) There are 62 millions of inhabitants in France in 2004.

Each of these values is more or less approximate. The direct answer is generated from the most precise numerical value if available (Moriceau, 2006). If all values are approximate, then the generated answer has to explain it: we plan to use prepositions of approximation (about, almost, ...) or linguistics clues which have been extracted from web pages (precision in the frames). The choice of a particular preposition depends on the degree of precision/approximation of numerical values.

4.2 Explanation Generation

Obviously, the generation of the cooperative part of the answer is the most complex because it requires complex lexical knowledge. We briefly present some of the necessary lexical resources. For example, verbs can be used in the answer to express numerical variations. For that purpose, we use the same classification of French verbs as for extraction, namely verbs of change and verbs of movement.
From these classes, we have characterized sub-classes of increase, decrease, etc., so that the lexicalisation task is constrained by the type of verbs which has to be used according to the variation mode (if verbs are extracted from web pages as linguistics clues of *variation*, they can also be reused in the answer).

A deep semantics of verbs (change, movement) is necessary to generate an answer which takes into account the characteristics of numerical variation as well as possible: for example, the variation mode, the speed and range of the variation. Thus, for each sub-class of verbs and its associated variation mode, we need a refined description of ontological domains and selectional restrictions so that an appropriate verb lexicalisation can be chosen.

For example, we use proportional series representing verb sub-classes according to the speed and amplitude of variation (cf. figure 5): the use of *climb* (resp. *drop*) indicates a faster growth (resp. decrease) than *go up* (resp. *go down*): the verb *climb* is preferred for the generation of *The increase of gas prices climb to 20.3% in October 2005* whereas *go up* is preferred in *The increase of gas prices go up to 7.2% in September 2005*.

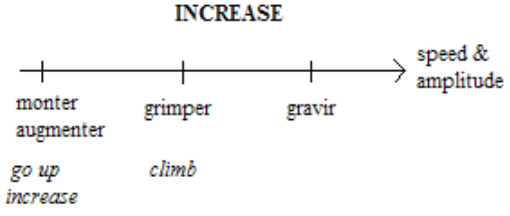

Figure 5 : Proportional series (increase)

As for direct answer generation, verbs can possibly be associated with a preposition that refines the information (*The average age of marriage increased by **about** 5.5 years between 1972 and 2005*).

9. Evaluation

Our system can select the correct direct answer provided that QRISTAL returns the correct answer among selected web pages and that our grammar succeeds in extracting relevant information. So, the extraction stage has to be evaluated according to 2 main points:

- evaluation of the quality of web pages selected by QRISTAL. Figure 6 presents some elements: we submitted 30 questions to Google and QRISTAL. QRISTAL returns the correct answer among relevant pages for 87% of the questions we evaluated,
- evaluation of our extraction grammar performances: does our grammar extract relevant information compared to what is manually extracted? This point is ongoing.

Google		
Correct Answer's Rank (average)	Incorrect Answer before the correct one	Relevant Pages (first 30 links)
4	43%	23%
QRISTAL		
Correct Answer	Correct Answer among relevant pages	Relevant Pages (first 30 links)
13%	87%	30%

Figure 6 : Elements of extraction evaluation

Concerning the variation characterization and generation evaluation, we have to evaluate:

- if the number of extracted information is sufficient to conclude that a numerical value varies and how it varies,
- if the generated direct answer is correct,
- if the generated explanation is comprehensible and considered as useful by the user.

10. Conclusion

In this paper, we presented an approach for the generation of cooperative numerical answers in a question-answering system. Our method allows us to generate:
(1) a correct synthetic answer over a whole set of data and,
(2) a cooperative part which explains the variation phenomenon to the user,
whenever several numerical values are extracted as possible answers to a question.
Information is first extracted from web pages so that numerical values can be characterized: variation criteria and mode are then identified in order to generate explanations to the user.

Besides evaluation, several future directions are obviously considered:

- an analysis of needs for common knowledge so that the answer characterization task is made easier,
- an analysis of how restrictions are lexicalized in texts (adjectives, relative clauses, etc.) in order to extract them easily,

- an evaluation of the knowledge costs and of what domain specific is (especially for common knowledge about restrictions),
- an evaluation of the quality of answers proposed to users and of the utility of a user model for the selection of the best answer.

10. References

Dahl, V. & Abramson, H. (1984). On Gapping Grammars. In Proceedings of the Second Logic Programming Conference, Uppsala, Sweden.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Originally published in London by Oliver and Boyd.

Harabagiu, S. & Lacatusu, F. 2004. Strategies for Advanced Question Answering. In Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004, Boston, USA.

Grice, H.-P. (1975). Logic and conversation. In P. Cole and J.L. Morgan, (eds.): Syntax and Semantics, Vol. 3, Speech Acts, New York, Academic Press.

Laurent, D. & Séguéla, P. (2005). QRISTAL, système de Questions-Réponses. In Proceedings of TALN, Dourdan, France.

Maurel, D. 1991. Préanalyse des adverbes de date du Français. TA Information, volume 32, n°2, p5-17, 1991.

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

Moriceau, V. & Saint-Dizier, P. (2003). A Conceptual Treatment of Metaphors for NLP. In Proceedings of ICON, Mysore, India.

Moriceau, V. (2006). Numerical Data Integration for Cooperative Question-Answering. In Proceedings of EACL-KRAQ, Trento, Italia.

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

Saint-Dizier, P. (1999). Alternations and Verb Semantic Classes for French. Predicative Forms for NL and LKB, P. Saint-Dizier (ed), Kluwer Academic.

Saint-Dizier, P. (2005). PrepNet: a Framework for Describing Prepositions: preliminary investigation results. In Proceedings of IWCS'05, Tilburg, The Netherlands.

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

Webber, B., Gardent, C. & Bos. J. 2002. Position statement: Inference in Question Answering. In Proceedings of LREC, Las Palmas, Spain.