A syntactic strategy for filtering sentences in a question answering system
Vincent Barbier and Anne-Laure Ligozat
LIMSI-CNRS
University of Paris Sud
BP 133, 91403 Orsay, France
barbier,annlor@limsi.fr

Abstract
In a question answering system, the first steps consist in retrieving documents relevant to the question, from which sentences are extracted. In these steps, the possible variations between the formulations of the question and the candidate sentences should be taken into account. The selection of documents has to be large enough to ensure a high recall, but the noise generated by the reformulations has to be contained. In this article, we will present a method for filtering and reranking the candidate sentences of the documents according to syntactic criteria.

1 Introduction
In a question answering system, the first steps consist in retrieving documents relevant to the question. This selection should ideally take into account the possible reformulations of the question, in order to ensure a high recall. But accepting important semantic variations leads to very noisy results, thus the documents retrieved have to be filtered. In order to filter them, a linear distance between the terms of the question found in the documents can be calculated. But this kind of distance is not very reliable. We chose to use instead a syntactic distance between these words in order to improve the precision of our selection.

In this paper, we will make a brief presentation of our question answering system QALC. Then we will detail the difficulties of passage selection, and the different strategies that can be used to face these difficulties. We will afterwards describe our solution, based on a syntactic filtering, and present an evaluation of this solution on a corpus of questions and answers. Finally, we will give some perspectives to our work.

2 Selection of relevant documents
2.1 QALC architecture
Our question answering system QALC is composed of four main modules: question analysis, document retrieval, document processing and answer extraction (Ferret et al. 02). The architecture of the system is described figure 1.

![Architecture of the QALC question answering system](image-url)

The question analysis module determines some information about the question: expected type of the answer, category of the question, keywords... This information is first used to retrieve documents thanks to the search engines, Lucene ¹ for a French corpus, and MG ² for an English one. These documents are then re-indexed by Fastr (Jacquemin 99) which recognizes morphological, syntactic and semantic variants of simple or composed terms of the question, and a subset of the highest ranked ones is kept. The named enti-

¹http://jakarta.apache.org/lucene/docs/index.html
ties tagging module is then applied to these documents. The final module is in charge of extracting the answers from weighted sentences: first, the sentences of the documents are weighted according to the presence of the terms of the question and of named entities and their linear distance, then, answers are extracted from the sentences, the process depending on the expected type of the answer.

2.2 Passage selection strategy

The sentences in which the answers are searched for are then the result of several successive selections:

- A first selection based on the non-empty words of the question retrieves the documents.
- Fastr proceeds to a second selection according to the recognition of mono and multi-term variants.
- Sentences are selected according to weights depending on the presence of the question terms and their linear distance, and on the presence of named entities of the expected type.

We chose to focus particularly on the third selection. The ranking of the sentences influences this of the answers, and thus it is crucial to be able to detect the sentences which are most likely to contain the answer. In order to assess the quality of our ranking of sentences, we calculated the Mean Reciprocal Rank (MRR) over the questions of the CLEF’04 multilingual question answering evaluation\(^4\) in which we participated. The MRR is of 0.306 for these questions. As an element of comparison, one can refer to (Tellex et al. 03) which made a comparison of several passage retrieval algorithms, and found MRRs ranging approximately from 0.26 and 0.43.

2.3 Selecting relevant passages for question answering

The output of the retrieval engine is usually reprocessed before extracting the answer, since this output is not best suited for question answering:

\[ MRR = \frac{1}{x} \sum_{i=1}^{x} \frac{1}{\text{rank}(\text{first} \ \text{answer})} \]

the documents may not be ranked if the engine is boolean, and their selection is driven by the keywords of the question rather than by the question itself. For example, the documents may not contain an entity of the expected type. This process can be partly mixed with the question extraction: most systems (re)rank the documents and restrict them to passages, and then process to the answer extraction using various strategies, ranging from expected type recognition to logic proof of the answer. The length of these passages can be more or less long.

For the passage selection, classic information retrieval models can be used, by relying on statistical information to evaluate the relevance of a document to a given query, for example with \( tf*idf \). But these methods are not completely adapted to question answering since short answer strings can be found in documents concerning completely different topics. Thus more question-driven strategies are required.

For instance, in the (Moldovan et al. 02) system, the passage selection module associates question terms with the set of their morphological alternations, and ranks the passages by estimating the degree of lexical matching between the question and the passages.

In (Hartrumpf 04)’s system, the selected passages are sentences; the sentences of the corpus of documents are transformed into semantic networks, and a semantic network matching the question is searched for.

In our system, we chose to process the answer extraction on sentence-long passages. As semantic reformulations were used by Fastr in the document selection, the sentence selection and ranking has to counterweight the loss of precision stemming from these reformulations. A deep semantic strategy was not chosen: first, it requires knowledge bases such as Extended WordNet constructed by LCC, which can hardly be used in a multilingual context, and robustness is difficultly achieved in a deep semantic system. Our hypothesis is that a syntactic filtering could also improve the passage selection.

3 Syntactic filtering

3.1 Sentence tree reduction

The ranking of the sentences according to a linear distance between the terms of the question presents drawbacks, since this distance does not
consider syntactic aspects. Sentence scores will be deteriorated by the presence of epithets or relatives, although these elements do not alter the meaning of the sentence. An example of this kind of problem is given Figure 2:

| Question: Who was married to Whoopi Goldberg? (Qui était marié avec Whoopi Goldberg ?) | Answer: Actress Whoopi Goldberg married film industry union representative Lyle Trachtenberg during a weekend ceremony at her Pacific Palisades home. |

Figure 2: An example question from the CLEF04 campaign

Another point justifying the use of a syntactic measure, is that it can favour sentences where the relevant words are closely linked to each other. Sentences where the question’s relevant words are not directly linked to each other should be less likely to answer the question.

We aim at creating a measure that takes the two above points into account.

Our solution consists in representing the sentences as syntactic graphs and pruning any phrase that is not useful to link the relevant terms together.

In order to be tested, this measure is inserted in the sentence weighting algorithm of the QALC system.

3.2 Algorithm

The algorithm prunes the syntactic tree of retrieved sentences so as to build the best subgraph containing the elements of the question, where “best” is defined through a measure combining syntactic and semantic proximity. In this approach, the syntactic structure of the question is not taken into account. The question is considered as a set of criteria, denoted $Q$.

3.2.1 Mapping between information of the question and the words of the answer.

The paradigmatic criteria for matching a sentence with the question are the following:

- The expected type of the answer.

For the moment, we only considered questions whose answer’s expected type are named or numerical entities. It seems possible to extend the algorithm to WordNet types without modifying the framework of the algorithm.

- The terms of the question

Terms of the questions can be lemmas, monoterms or multi-terms, verbs, nouns or noun phrases. They are linked to either monoterms or multi-terms in the answer. There can be semantic or morphological variations between the terms of the question and those of the answer. Words that appear in both answer and question without variation are referred to as lemmas.

The paradigmatic links are weighted: named entities, Fastr terms and lemmas of the question term are given different weights. For example, named entities have a weight of 2. Fastr term’s weight vary according to the reliability of the term: For example, a bi-term should be scored higher than a mono-term.

3.2.2 Selecting the best combination of nodes

For each element of the question, either a term or the expected answer type, we obtain the list of the nodes of the answer’s graph that are likely to be paradigmatically linked to this element.

If the element is a term of the question, it may point to a composed term of the answer. In that case it corresponds to more than one node, but will be treated as a simple term that will correspond to the head of the composed term.

In these lists of nodes there is bound to be non-relevant elements. We try to filter the non-relevant nodes by selecting a combination of nodes, keeping only one node for each criterion, so we can denote $weight(c)$ the weight of the paradigmatic link selected for criterion $c$. We chose the combination that maximizes a weight combining a syntagmatic and a paradigmatic constraint. The weight of a combination is computed as follows:

**Syntagmatic part of the weight:** We build the minimal subtree containing all the nodes of the combination, as shown in figure 3.

Let $nb\_nodes$ the total number of nodes of this graph and $nb\_criteria$ the number of nodes...
linked to a criterion of the question. Note that $nb_{criteria} < nb_{nodes}$. We calculate a density which is given by the function shown in figure 4.

The function represents a fuzzy threshold function. The threshold limit is set to $nb_{nodes} = 2 \times nb_{criteria} + 1$, which corresponds to the case of an answer graph where there is always exactly one non-relevant node between two relevant nodes.

**Paradigmatic part of the weight:** In order to put a disadvantage on the nodes which are not strongly related to the question’s element, we take the weights of the paradigmatic relations into account. The final measure is:

$$\text{weight(combination)} = \text{density} \times \sum_{c \in Q} \text{weight}(c)$$

### 3.2.3 Connecting the metric into the QALC architecture

The QALC system performs a sentence ranking which integrates several measures such as answer terms linear distance, Fastr terms weights, named entity weights into a global sentence weight. We connect our syntactic measure by multiplying this global weight by the density we computed. The sum of weights used for determining the best combination is no longer used in this step, for it is redundant with the QALC systems weights.

### 3.3 Study of our approach on the Clef 04 corpus

We evaluated our strategy over the CLEF04 corpus of questions. It has to be noted that CLEF being a multilingual evaluation, our evaluation on this corpus suffered sometimes from translation term difficulties. Table 1 shows the results of this evaluation.

<table>
<thead>
<tr>
<th></th>
<th>NE questions</th>
<th>All questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial MRR</td>
<td>0.310</td>
<td>0.306</td>
</tr>
<tr>
<td>New MRR</td>
<td>0.360</td>
<td>0.338</td>
</tr>
</tbody>
</table>

Table 1: MRR with and without syntactic reranking

The most significant figures are those concerning NE questions, since our reranking of sentences is presently restricted to those, but the overall improvement is nevertheless interesting since the MRR also increases significantly, due to the relatively high percentage of NE questions. On NE questions, we improved our MRR by 17%.

To the question “En quelle année le Pape Jean Paul II est-il devenu pontife ?” (“In which year did Pope John Paul II become pontiff?”) the following sentence gained 20 positions for example:

“(...)but never again will the election of a non-Italian Pope be as startling as when Cardinal Karol Wojtyla of Krakow was elected in 1978.”

The minimal subgraph containing “Pope” and a DATE named entity can indeed be represented as shown in Figure 5, bold font words are those contained by the pruned structure.

Another example of reranking is given by question “Qui a gagné le Prix Nobel de Littérature en 1994 ?” (“Who won the Nobel Prize for literature in 1994?”), for which the answer “Derek Walcott, who won the Nobel Prize for litterature, called (...)” is ranked 7 instead of 18, thanks to the subgraph also shown in Figure 5. Note that the “1994” criterion is absent from the sentence but is present in the retrieved document, thanks to the search engine selection.

### 3.4 Perspectives

As noticed earlier, the syntactic structure could be taken into account, in order to privilege in the minimal subtree construction, the links between terms which were highly related in the question. Moreover, the strategy could be extended to non-NE
questions, in case the semantic type of the answer can be found and verified.

Another improvement would be the use of tree edit distances to approximate syntactic similarity. (Kouleykov & Magnini 05)

This strategy could also benefit from other types of reformulations; WordNet variants could also be considered as question term reformulations. Finally, it would be interesting to test this sentence filtering one step before in our system, and to compare fully and combine the selections based on a linear distance and on a syntactic one.

4 Conclusion

By using a syntactic distance instead of a linear one to select sentences in our question answering system, we improved the ranking of these sentences, and thus our probability to find the correct answers. The type of questions for which this strategy is most relevant could be studied, in order to try and detect to which extent this strategy can replace our previous one.

References


