Les Cahiers du GERAD ISSN: 0711–2440

Performance of n-Grams for a Question Retrieval System in the Context of Approximated Spelling

E.-S. Tang, F. Bellavance,
N. Péladeau, G. Caporossi

G–2008–60

September 2008

La publication de ces rapports de recherche bénéficie d'une subvention du Fonds québécois de la recherche sur la nature et les technologies.

Performance of n-Grams for a Question Retrieval System in the Context of Approximated Spelling

Eng-Seng Tang

HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7

François Bellavance

GERAD and HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7
francois.bellavance@hec.ca

Normand Péladeau

Provalis Research
2414, avenue Bennett
Montréal (Québec) Canada, H3T 2A7

Gilles Caporossi

GERAD and HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7
gilles.caporossi@hec.ca

September 2008

Les Cahiers du GERAD
G–2008–60

Abstract

Question retrieval systems, unlike question answering systems, exploit the knowledge contained in previously answered questions to answer new ones by returning already answered question that may respond to the user's information needs. In our experiment, we work on improving a French language question retrieval addressing mostly young people's questions with approximate spelling. To assess the spelling problem, character n-gram features have been proposed in the literature as an alternative to the classical word based features. In the present study, we compare the performances of question retrieval models using character n-grams (for n = 3, 4 and 5) to ones obtained using the classical baseline word based features. Furthermore, we test the "simplified French" procedure which attempt to improve the performance of the model by simplifying the French writing. Our results show that if n-grams do not perform as well as we could expect but perform rather well in the case of 4-gram together with the simplified French procedure.

Résumé

Contrairement aux systèmes de réponse automatiques, les systèmes d'extraction de questions exploitent les informations contenues dans les questions auxquelles on a déjà répondu en fournissant à l'usager des questions auxquelles on a déjà répondu qui peuvent les satisfaire. Dans notre expérience, nous travaillons à l'amélioration d'un système d'extraction de questions destiné principalement à des jeunes, dans un contexte avec orthographe approximative. Pour répondre à ce problème, les n-grams ont été proposés dans la littérature comme une alternative aux caractéristiques de textes basées sur les mots. Dans cette étude, nous comparons les performances de modèles d'extraction de questions utilisant les n-grams (avec n = 3, 4 et 5) à ceux utilisant des mots. De plus, nous testons une procédure de français simplifié qui tente d'améliorer la performance des modèles en simplifiant l'écriture du français. Nos résultats montrent que si les n-grams ne performent pas aussi bien qu'excompté, mais performent relativement bien dans le cas des 4-grams s'ils sont associés à la procédure de français simplifié.

Introduction

In the present paper, we present a comparison of different models for Question and Answer retrieval system on a real application in a context of approximative spelling. The data used is that of *Tel-Jeunes*, a free social intervention service dedicated to the help of young people with their personal difficulties. This "real life" data bank has three special characteristics:

- the number of misspelling errors is very important and some techniques to handle misspelling errors could fail when they are so numerous.
- The questions asked by the users of the system are not factual at all. Most of the time, a question is associated to a long description of the context.
- Due to the human dimension of the task achieved by the system, the range of questions to be handled is very large and one cannot be sure that the question bank, even if it is large, contains at least one satisfactory answer to a given question.

In this study, we propose a comparison of models based upon complete words, 3-grams, 4-grams or 5-grams. In addition, each of these models could be applied to raw texts or texts converted by the "simplified French" procedure, a procedure that aims at reducing the impact of spelling errors by simplifying the writing. The algorithm used for question retrieval is the K-nearest neighbors as it is known to perform properly for this kind of task [13].

The paper is organized as follows : the first section of the paper describes the context, the *Tel-Jeunes* question answering system, the experimental setting is described in the second section and the results are described in the third section. The last section finally exposes a short discussion and conclusion.

1 The Tel-Jeunes question aswering system

Located in the province of Quebec, in Canada, Tel-Jeunes is a free social intervention service that is dedicated to the help of young people from age 5 through 20 with their personal difficulties. Within the service, professional social workers give assistance, support and answer questions of people in need of help. At first, the service was exclusively given through a toll free phone line but since February 2001 an on-line service was created and allowed the users to write questions to the social workers and receive an answer within 24 hours. A virtual mailbox was created to receive the users' questions and all the received inquiry as well as the returned answers were stored to build a valuable question bank. Within this bank, each questions-answer (Q&A) pair is categorized by the social workers into one of a variety of pre-defined semantic/topic categories. Due to a high volume of questions being asked through the on-line service and the limited number of written questions that can be taken by the social workers, only a restrained number of messages could be accepted daily in the mailbox. Everyday, that mailbox would fill up very quickly and would be available to the young users only 27% of the time. Consequently, a lot of them were simply not given the opportunity to write questions to Tel-Jeunes. In response to that problem and to leverage the fact that many of the asked questions are similar in nature, Tel-Jeunes recently implemented a question retrieval system on their website where younglings can write questions in and have the system search for relevant Q&A pairs in the built question bank Doing so, the objective of the system is to assist Tel-Jeunes in helping more users by eliminating the need for social workers to process the questions that they have already answered in the past. In the latter case, the system

also has the advantage of giving immediate answers to the users and therefore improves the service offered by Tel-Jeunes.

1.1 Overview of the Q&A system

Basically, the Tel-Jeunes' system behaves as many Frequently Asked Questions (FAQs) retrieval systems but also has some very particular aspects but it operates in a French environment and the questions to treat usually yields a huge number of misspelling. For these reasons, it is difficult the use of any semantical information and the system must rely on the statistical information from the term frequencies to achieve its task.

To each Q&A pair from the database is first associated a vector representing the Question part using term frequencies. From these vectors are then computed distances to the new Question. The so computed distances are lately used to select the appropriate Q&A pairs to show the user as illustrated on Figure 1.

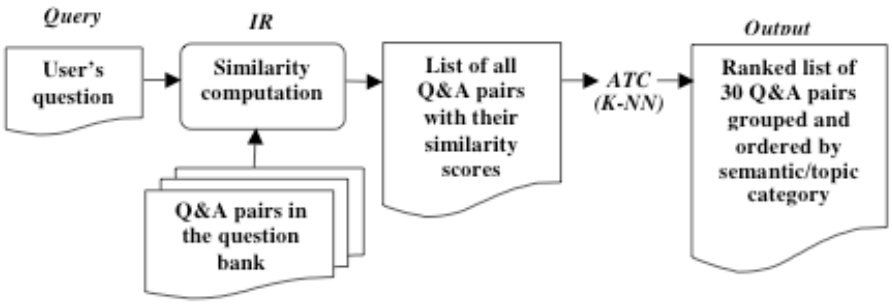

Figure 1: Representation of the Q&A system.

1.2 The information retrieval (IR) module

The question retrieval system built for Tel-Jeunes is an IR system based on a vector space model (VSM) [2]. This model is build using the following steps: *WordStat* applies a series of transformations and encoding to the text to be able to treat it properly. The first part aims at unifying all texts, then some procedures are used to access the misspelling problem and to reduce the influence of useless words (for the purpose of the present task). Then it uses n-grams to build the model.

1.2.1 Preprocessing

Before any further operation, *WordStat* standardize the text by up-casing all letters and removing all special characters such as punctuation symbols. Then, a series of transformations are applied to the text in order to improve the reliability of the system as described in the following sections.

1.2.2 The extraction of n-grams

The n-grams extraction procedure considers n consecutive letters instead of the word itself. For example, the word *premier* would be encoded as the following 3-grams: *p, pr, pre, rem,*

emi, mie, ier, er and *r*. The procedure used by *WordStat* considers words boundaries and introduces spaces before and after each word to avoid overlap between words. This procedure tends to reduce the impact of slight typos as most of the n-grams involved will not be affected by the error [1] [4].

1.2.3 Assessing the approximate spelling problem, the simplified French procedure

In the Tel-Jeunes database, a lot of misspelling errors were identified, mainly related to phonetic writing. To be efficient, the system must thus be robust enough to perform properly even if the texts contain lots of spelling errors.

The simplified French (SF) procedure rewrites each word in a simplified way by undoubling all consecutive double letters and removing all accents. The simplified French procedure could be used prior to the extraction of n-grams. For example, the words *appeler, appeller* and *apeler* are all encoded as *apeler*, and the words *accèder* and *accéder* are both encoded *aceder*.

1.2.4 Term selection and weighting

The goal of the system is to find in the database of answered questions the closest ones so that the answer given is satisfactory. In this case, emphasis must be given on the topic rather than the way the question is asked. The "grammatical" words are completely useless in this task and should not be considered. Removing the more frequent words achieves this task. On the other hand, words that are very rare cannot really help either and should be removed too. Frequent and infrequent terms removal strategies have been well documented in early works on IR done by Salton and McGill (1983). For our experiment, terms that occurred in 90% or more of the documents have been removed because they are not useful in differentiating between documents. Secondly, for efficiency reasons, we have decided to also eliminate terms which have a frequency inferior to 8. According to Salton and McGill [9], these are so sparsely found that they should not significantly affect the retrieval performances of the models. There were no specific guidelines in the literature regarding the choice of these values thus they were chosen arbitrarily. However, these are probably conservative values because very few terms would meet those criteria and be removed.

Even if some words are removed because they negatively affect the performance of the model, all the remaining terms will not be of the same level of interest; therefore, a weighting strategy for the terms used will certainly increase the performance of the system. In our experiment, we will use the well established $TF-IDF$ (Term Frequency - Inverse Document Frequency) scheme. With $TF-IDF$, each term is given a weight for each document as described in equation 1:

$$w_{tj} = TF_{tj} \times IDF_t. \tag{1}$$

There are different versions of the $TF-IDF$ scheme in the literature, but the most common ones uses the TF definition from equations (2) and (3) and IDF as defined in equation (4).

$$TF_{tj} = freq_{tj} \tag{2}$$

$$TF_{tj} = \frac{freq_{tj}}{\max_l freq_{lj}} \tag{3}$$

$$IDF_t = \log \frac{N}{n_t}, \tag{4}$$

where $freq_{tj}$ represents the frequency of the term t in the document j, $\max lfreq_{lj}$ denotes the largest of all the frequencies of the terms in document j, N is the total number of documents in the collection and n_t is the number of documents containing the term t.

1.2.5 Computing similarities among documents

After these treatments, each the text (document) j is represented as values in the vector space of the terms $t = 1 \dots T$ as described in equation (5),

$$\vec{D_j} = (w_{1j}, w_{2j}, \dots, w_{Tj}). \tag{5}$$

Using this vector representation of texts, there are different ways to compute similarities or dissimilarities between texts. The easiest would be

$$diss(d_i, d_j) = \sum_{t=0}^{T} |w_{ti} - w_{tj}|. \tag{6}$$

This formula computes a dissimilarity value based upon the Manhattan distance. If the weight w_{ti} of the term t in document i is a binary value (indicating the presence/absence of the term in the document), this measure is known as hamming distance.

Another measure defined upon binary variables is the Jaccard given by the Jaccard coefficient taking asymmetry in the data into account as follows:

$$sim(d_i, d_j) = \frac{\sum_{t=1}^{T} w_{ti} \times w_{tj}}{\sum_{t=1}^{T} \max(w_{ti}, w_{tj})}. \tag{7}$$

Note that the Jaccard coefficient is a similitude coefficient taking larger values when the documents are more similar (it is the opposite for distance measures).

In the present work, we use cosinus of the angle θ between the two vectors $\vec{D_i}$ and $\vec{D_j}$ as measure of similitude as given on equation (8).

$$sim(d_i, d_j) = \cos(\theta) = \frac{\vec{D_i}.\vec{Dj}}{||\vec{D_i}||.||\vec{D_j}||} \tag{8}$$

If the weights w_{ti} are all non negatives, $0 \leq \cos(\theta) \leq 1$. This methods gives then a measure of partial similarity between the request and the document that has the advantage of using a more precise information than the measures based upon binary variables. Compared to the raw Manhattan distance, it provides a more accurate measure as it involves an implicit scaling.

1.3 The automated text categorization (ATC)

Various methods may be considered for the text categorization task such as support vector machines (SVM) or artificial neural networks (ANN) for example. We suggest to the reader willing to learn more about these various methods to read the survey by Sebastiani [10]. In the present application, we will use the k-nearest neighbors algorithm (K-NN). Using a set of already classified documents, the K-NN algorithm first finds the k closest ones and classifies the new document (the question in our application) according to these k documents. Three parameters must be chosen to use this approach:

Table 1: Distribution of the questions of the 1st data set within the twelve semantic categories

Category	Number of questions	Percentage
Health	161	2.1
Intimidation	124	1.6
Violence	99	1.3
Sexuality	2635	33.7
Love & Relations	2136	27.3
Personal problems	556	7.1
Rights & Law	92	1.2
Gambling	4	0.1
School	139	1.8
Drugs	201	2.6
Pregnancy/Abortion	552	7.1
Friends/Family	1122	14.3
Total	7821	100

1. The number of neighbors to use k,
2. the distance or dissimilarity measure,
3. and the rules to assign the new document a class.

2 Experiments

In this section, we will describe the methodology used to define the best choices for various parameters or strategies involved according to our problem and the data used.

2.1 Data description

Tel-Jeunes provided us two question sets for the experiment. The first set is composed of 7821 previously answered questions asked by users between January 1st of 2003 and March 22nd of 2006 with the exception of questions that are deemed too sensitive, like questions about suicide. Tel-Jeunes considers that young people with these types of questions should be looking for an access to the social workers directly so that they can be help in the most appropriate way.

All the pre-answered questions include the text of the questions as well as their titles as given by the users but not their answer parts, which are not used in anyways in the retrieval process. No spelling corrections or modifications have been applied to these questions meaning that they were kept the way they were written by the users. As stated, all questions were manually categorized by Tel-Jeunes' social workers into pre-defined categories. Twelve categories can be found in the provided sets. These are the coarse level categories used by Tel-Jeunes. The social workers actually also classify the questions into finer grain categories but these were not provided. Table 1 lists the number and percentage of questions found in each category for the first data set.

In average, at least one word out of 8 written in the questions is misspelled and this rate grows to one word out of 4 in about 10% of the questions. For example, we found 43

variation of the word *anorexique (anorexic)* such as *annorexique* or even *anorexik*. In addition to misspelling, some of the words simply do not exist in French, which is due to aggregation or abbreviations of words. For example, *j'vais* or *jvais* is used instead of *je vais (I will)*; these errors can certainly be attributed to a phonetic style of writting.

One distinctive aspect of the questions written to social workers is that they usually are not factoid and may often rather be considered as open discussion. This results in complex multi-topic questions wrapped in between words that may not be relevant to the questions themselves.

As an indication, the average length of a question is close to 600 characters while the questions in the question answering tasks of the TREC are only a few words long [12].

2.2 Building and training the question retrieval (QR) models

Using the *WordStat's* classifier module and the questions from the first question set, the objective of this phase was to build the most performant QR models for each feature type. Note that *WordStat* does not directly create a QR model but allows us to build a K-NN classifier based on an IR vector space model. To obtain the QR models, we simply use the classifiers to find the most similar questions to an input question. It is important to note that for the experiment, we did not reproduce in the QR models the last step that is done by the Tel-Jeunes system in which the returned questions are grouped by semantic category. In that sense, the built QR models work like traditional IR systems and simply return a ranked list of relevant questions.

To test the QR models we used a leave-one-out cross validation setup. This setup consists of using all the questions of the set except for one to build a QR model, to use the left out question as an input question and to assess the relevance of the questions returned by the model relatively to that input question. This process is then repeated for all the questions of the set, each time a new question playing the role of the left out question.

However, at the time of experiment, the relevancy between each of the questions of the provided set was not known. Therefore, it was impossible for us to build and optimize the different QR models based on their retrieval performance; we had no way of assessing the relevance of the returned questions. Theoretically, we however made the assumption that the relevance of the suggestions returned by the models to an input question is highly category dependant: if an input question is affected to the right category then the questions within that category have a higher probably of being relevant to it than questions from other categories. Since we had in hand the categories of the questions, we instead optimized the QR models based on the categorization performances of the classifier from which they are obtained.

The question categorization problem is a classical single label, multi-class one. Thus, for optimizing the QR models during the construction phase, the most appropriate measure of effectiveness is the *accuracy* measure [10] which represents the proportion of correctly classified questions. The measure is computed as follow:

$$A = \frac{n_c}{N} \tag{9}$$

where A is the accuracy, n_c the number of correctly classified questions and N is the total number of questions.

For building the classifiers, *WordStat* requires the selection of four parameters:

- The K parameter value of the K-NN algorithm;
- The metric for feature selection;
- The metric for term weighting;
- The number of features to consider for each QR model.

Here we detail the way the values for those parameters were determined.

2.2.1 K-parameter value of the K-NN algorithm

The K-parameter value of the K-NN algorithm determines the number of Q&A pairs returned to the user. The Tel-Jeunes actual system returns 30 Q&A pairs, which is rather conservative and allows the user to access a large number of Q&A pairs (even if most users do not look at all of them). However, for our experiments, we used a value of 10, motivated by our beliefs that most users are not likely to look past the first few returned Q&A pairs.

2.2.2 Determining the feature selection and term weighting metrics

WordStat offers multiple feature selection and term weighting metrics to choose from. The feature selection could be based on the global Chi-Square or the Max Chi-Square computed on the document occurrence, the term frequency or the percentage of terms. 6 combinations of feature selection are thus available. The term weighting could be based upon the document occurrence, the term frequency or the percentage of terms that could be used as such or weighted by inverse document frequency, Chi-Square or Max Chi-Square. Overall, 72 combinations of feature selection and term weightings are available in *WordStat*.

To determine which combination of statistics to consider for each model, we did extensive experimentations on the categorization performance that could be obtained from models using them with a number of features from 50 to 55000 (only 3923 in the case of 3-gram as it is the total number of different 3-grams found in our bank).

For these tests, the K parameter value of the K-NN algorithm was set to 30; the same value used in the Tel-Jeunes QR system. From the results, we determined that the best overall categorization results were obtained by using, for every model, the *Max Chi-Square* statistic [5] computed on the *Percentage of terms* (which is the relative frequency of a term in the document) for the feature selection metrics coupled with term weighting based on the *TF-IDF* scheme defined by equations (2) and (4).

2.2.3 Number of features to considerer

The last parameter to determine for building the QR models is the number of features to consider. Using the *Max Chi-Square* statistic computed on the *Percentage of terms* for feature selection, a term weighting scheme based on *Term Frequency* and weighted by the *Inverse document frequency* and a K parameter value of 10 for the K-NN algorithm, we undertook the last categorization experiments where we tested QR models for each of the four feature types, with and without the simplified French procedure, for a number of features varying from 100 to 2600 with a step of 100. The decision to limit the number of features at 2600 was motivated by the test results obtained when determining the feature selection and term weighting metrics. In those tests, the accuracy of the different models seemed to stagnate

Table 2: Number of features to reach highest accuracy for each QR model.

Model	Highest Accuracy	Nb Features
Word	0.84	1900
3-gram	0.81	1900
4-gram	0.82	2600
5-gram	0.81	2600
Word-SF	0.84	1700
3-gram-SF	0.80	1600
4-gram-SF	0.82	2600
5-gram-SF	0.81	2600

around that number. To choose the number of features for each QR model, we looked at the highest accuracy that is obtained by each of them and retained the number of features with which they obtained that result. Table 2 presents the highest accuracy obtained by each model and the respective number of features used to reach that accuracy.

The best categorization performance was obtained by the word based model. This result may seem surprising since we expected the n-gram based models to perform better than the word based one. Also note that the 4-grams and 5-grams models obtained their highest accuracy with the maximum number of features allowed.

To summarize, the feature selection and term weighting metrics are the same for all models as well as a value of 10 for the K parameter of the K-NN algorithm. The number of features for each QR model, however, is the one with which it obtained the highest accuracy.

3 Results

To assess the effectiveness of the QR models, we confronted each of them to a set of 150 unanswered questions randomly chosen and had them retrieve 10 previously answered questions from the 7983 questions bank only using the question part of the Q&A pairs. To avoid confusion between the input questions and the retrieved questions, we will refer to the latter as suggestions. Note that it was not verified if relevant suggestions to the 150 chosen questions could be found in the 7821 questions of the first set.

The performance of an IR system is usually assessed by classic measures such as precision, recall and F1-score (Lewis, 1991). However, the inter-relevancy of the questions in the question bank is not known, therefore making it impossible to use these measures. We instead asked two social workers from Tel-Jeunes to rate the relevancy of the 10 suggestions returned by each of the eight QR models relatively to the 150 selected questions.

The suggestions were given randomly to each social worker so they couldn't determine which QR model provided them. The number of suggestions to be rated by the social workers was between 1500 and 12000 depending on the consistency of the various models (number of different questions/suggestions pairs according to different models used).

To allow us to assess if the two social workers rated the suggestions in the same manner, i.e., the inter-rater agreement, the suggestions returned to 20 of the 150 questions were rated

Table 3: Contingency table of the relevancy scores given by the social worker 1 (columns) and 2 (lines) to the suggestions returned by the QR models.

	0	1	2	3	Total
0	85.6%	2.2%			87.8%
1	3.3%	2.5%			5.9%
2	2%	1.8%	1.2%	0.2%	5.2%
3		1.2%			1.2%
Total	91%	7.7%	1.2%	0.2%	100%

by both social workers. Before rating the suggestions, the social workers indicated, via a four point Likert scale of zero to three, zero being *not likely* and three being *very likely,* the likeliness of finding a relevant suggestion to the input question in the Tel-Jeunes question bank. If a score of zero is attributed to a question then it is simply eliminated from the evaluation phase. Otherwise, the suggestions returned by each model to the input questions are rated, on another four point Likert scale of zero to three, zero being *not relevant* and three being *very relevant*, by the social workers on their degree of relevance to the input question.

The suggestions returned by the model to 20 questions out of the 150 chosen were rated by both social workers. 598 suggestions were rated by the social workers relatively to those 20 questions (two of them were missed by the workers). Before any further analysis of the results, the inter-rated agreement was verified in order to validate the quality of the rating. Table 3 is a contingency table of the relevance scores given by each of the social workers to the suggestions relatively to the input questions. The results suggest that rater 1 is more firm than rater 2 when evaluating the relevancy of the suggestions as there is 8.3% of the suggestions for which the former gave a lower score than the latter. Only 2.4% of the suggestions received a lower score from rater 2 comparatively to rater 1. Overall, the raters perfectly agreed on the relevance of 89.3% of the suggestions. Out of this number, 85.6% of the agreement cases were when both gave the score 0 to the suggestion. Furthermore, in 96.8% of the cases, the rating given by the social workers only differed by one point. The weighted Kappa agreement coefficient is 0.502 (95% C.I.: 0.412 – 0.593) which indicates a moderate but sufficient agreement for the purpose of our study.

3.1 Performance measure

To avoid bias that may occur when using a performance measure, four measures were used to evaluate and compare the retrieval performances of the QR models.

1. The mean relevance score (MRS) of the suggestions returned by each model relatively to the input questions,
2. the proportion of questions for which the models returned at least one suggestion that was deemed to be relevant by the social workers (PR),
3. the mean reciprocal rank (MRR) and the weighted MRR (W-MRR),
4. and the alternative score (AS), related to the W-MRR.

The first two measures are very intuitive in their meaning but only consider a portion of the information in a different way. With the MRS, scores are averaged, which means that

Table 4: Distribution of the suggestions relevancy score and mean relevancy score by QR model.

Scores	**0**	**1**	**2**	**3**	**Mean**
Words	81.1%	9.7%	5.4%	3.8%	0.318
3-grams	81.4%	9.7%	5.2%	3.7%	0.311
4-grams	80.3%	10.8%	5.2%	3.7%	0.323
5-grams	82.6%	10.1%	4.0%	3.3%	0.280
Words SF	79.8%	10.8%	6.3%	3.2%	0.329
3-grams SF	83.4%	8.3%	4.8%	3.5%	0.285
4-grams SF	78.7%	11.2%	5.6%	4.5%	0.360
5-grams SF	81.4%	9.6%	5.1%	3.9%	0.315
Total	**81.1%**	**10.8%**	**5.2%**	**3.7%**	**0.323**

a model providing a good suggestion at first rank would be evaluated in the same way as a model which provides many suggestions that are a little relevant. On the opposite, the second measure only considers the proportion of question for which at least a relevant suggestion is provided by the model. The two last measures are build to consider both aspects and will be described in details.

The models performance were computed using the 10 first suggestions they provided for each question.

3.1.1 Mean relevancy score of returned suggestions

Table 4 presents the distribution of the suggestions' relevancy scores as well as the mean relevancy score of the suggestions returned by each QR model. When considering all models, most of the returned suggestions (81.1%) received the relevancy score of 0 meaning that most would not meet the information needs of the users who asked the questions and only a small fraction of the suggestions (3.7%) were deemed very relevant by the social workers. This results are not surprising since we don't know wether such a suggestion is available in the bank. Therefore, it seems difficult for the models to return very relevant suggestions.

Wether the simplified French procedure is used or not, the 4-gram based model obtains the highest mean relevancy score and the highest proportion of suggestions judged to be at least a little relevant. Interestingly, the word based model achieved a higher mean relevancy score than both the 3-gram and 5-gram based models. The simplified French procedure usually improves the quality of the models (except the 3-gram based model for which it drops from 0.311 to 0.285).

Overall, the higher mean relevancy score was obtained by the 4-gram based model with the simplified French procedure applied.

Table 5: Distribution of the maximum relevancy score of the suggestions returned to each question by QR model.

Scores	0	1	2	3
Words	27.1%	34.0%	18.1%	20.8%
3-grams	31.9%	34.0%	13.9%	20.1%
4-grams	31.9%	30.6%	16.7%	20.8%
5-grams	39.6%	30.6%	11.8%	18.1%
Words SF	22.9%	34.0%	22.2%	20.8%
3-grams SF	38.9%	28.5%	14.6%	18.1%
4-grams SF	31.9%	27.8%	18.1%	22.2%
5-grams SF	35.4%	27.8%	16.7%	20.1%

3.1.2 Proportion of questions for which there is at least one relevant suggestion returned.

If we suppose that, from the perspective of a user, only one relevant or very relevant suggestion in the returned list is enough to answer his question then this measure can be perceived as the proportion of questions adequately answered by each model.

For each of the 144 questions, the maximum relevancy score among the 10 suggestions returned by each model is considered. Table 5 presents the distribution of these maximum relevancy scores by QR model, which represents the aptitude of the model to provide at least one relevant suggestion to the user.

We first suppose that a suggestion with a relevancy score of at least 2 (relevant or very relevant) is relevant enough for answering the submitted question. When the simplified French procedure is not applied, the word based model that obtains the highest score answering 38.9% of the submitted questions with the 4-gram based model trailing not far behind answering 37.5% of the questions. When the simplified French procedure is applied, the word and the 4-gram based models are again respectively the best and second best model with 43% and 40.3% of the questions answered.

Overall, if we consider that suggestions with relevancy of 2 or 3 are enough to answer the input question then the best model is the word based one combined with the simplified French procedure. However, if we consider that a relevancy score of 3 is needed to answer the questions the model using 4-grams, with the simplified French procedure performs the best.

3.1.3 Weighted mean reciprocal rank and weighted mean alternative scores

To take in account the ranking of the suggestions, the mean reciprocal rank (MRR) measure that was used in question answering tasks of TREC [11] and of the national institute of informatics test collection for IR systems (NTCIR) workshop [6]. The reciprocal rank (RR) measure supposes that relevant suggestions are more valuable when they appear at the beginning of the ranked list than when they appear further down, i.e., it measures how "fast" a model is able to return a relevant suggestion to the user. This measure only considers the first relevant suggestion returned to an input question and doesn't give any credit if other relevant suggestions are also returned in the list. For a given question i submitted to the system, the

Table 6: Weighted mean reciprocal rank score obtained by each model according to the minimum relevancy required to consider a suggestion valid.

Weighted MRR	**1, 2 or 3**	**2 or 3**	**3**
Words	0.100	0.068	0.038
3-grams	0.103	0.074	0.038
4-grams	0.108	0.074	0.039
5-grams	0.091	0.060	0.032
Words SF	0.097	0.066	0.030
3-grams SF	0.097	0.071	0.036
4-grams SF	0.115	0.080	0.042
5-gams SF	0.098	0.069	0.039

score is $RR_i = \frac{1}{r_i}$ where r_i is the rank of the first relevant suggestion. If no relevant suggestion is returned, $RR_i = 0$. The MRR is the mean of the reciprocal rank scores across all the n questions submitted to a model as given by equation (10)

$$MRR = \frac{\sum_{i=1}^{n} \frac{1}{r_i}}{n}. \tag{10}$$

To take the degree of relevance of a question into account, which is specific to our application, we define the Weighted MRR, based upon the MRR measure given by equation (11)

$$W - MRR = \frac{\sum_{i=1}^{n} RS_i \times RR_i}{n}, \tag{11}$$

where RS_i is the relevance score given by the social workers to the first relevant suggestion returned by the model to question i.

A close look at the MRR score suggest that it does mainly take the best ranks into account. Indeed, dropping a suggestion from the first rank to the second decreases its RR by $\frac{1}{2}$, which is large compared to the corresponding performance decrease in practice. To assess this problem, we defined a new score, called Alternative Score (AS) as described in equation (12).

$$AS_i = 1 - \frac{r_i - 1}{k}, \tag{12}$$

where k is the number of suggestions considered (10 in the current study).

Word based models do not seem to perform as well on the weighted MRR and weighted MAS measures. Some of the trends observed in the previous measures are however reflected in the results obtained for these performance measures. Tables 6 and 7 respectively present the score for each QR model on the weighted MRR and weighed MAS measures. Since the question about how relevant a suggestion should be considered relevant by a user is debatable, we've included the scores for three different situations. We first considered that suggestions with relevancy scores of 1, 2 or 3 are relevant to a user, then only suggestions with score 2 or 3 and lastly, only suggestions with a score of 3.

When the simplified French procedure is not applied, the 4-gram based model, with a score of 0.108, is the top performing model in the situation where suggestions with relevancy scores

Table 7: Weighted mean alternative score obtained by each model according to the minimum relevancy required to consider a suggestion valid.

Weighted MAS	**1, 2 or 3**	**2 or 3**	**3**
Words	0.183	0.126	0.069
3-grams	0.186	0.132	0.066
4-grams	0.197	0.134	0.074
5-grams	0.168	0.108	0.062
Words SF	0.187	0.127	0.057
3-grams SF	0.172	0.124	0.063
4-grams SF	0.211	0.146	0.080
5-gams SF	0.182	0.127	0.072

Table 8: Performance variation when applying the simplified French procedure to different models.

Evaluation measure	**Word**	**3-gram**	**4-gram**	**5-gram**
M-Rel	+ 3.46%	- 8.36%	+ 11.46%	+ 12.50%
Prop. 2-3	+ 10.54%	- 3.82%	+ 7.47%	+ 23.10%
Weighted MRR	- 3.00%	- 5.83%	+ 6.48%	+ 7.69%
Weighed MAS	+ 2.19%	- 7.53%	+ 7.11%	+ 8.33%

of 1, 2 and 3 are considered relevant. However, the 3-gram and word based models perform very similarly to the 4-gram based one when only suggestions with relevance scores of 2 and 3 or when only suggestions with a score of 3 are considered relevant. It is difficult to say which model performs best. The 5-gram based model is clearly the worst one always obtaining the lowest performance score.

When the simplified French procedure is applied, the 4-gram model is clearly the most performant model with scores of 0.115, 0.080 and 0.042.

When assessing the performance of the models with the weighted MAS measure, it is the 4-gram based model that performs the best wether the simplified French procedure is used or not.

Overall, on the weighted MRR and weighed MAS measures, the best scores are obtained once again by the model using 4-grams with the simplified French procedure applied.

3.2 Impact of the simplified French procedure

Table 8 presents the performance variation on all 4 evaluation measures (mean relevancy, proportion of question for which there is at least one suggestion of score 2 or 3, weighted MRR and weighted MAS) of the different models when the simplified French procedure is applied compared to when it is not applied. The simplified French procedures seems to improve the models for most of the evaluations measures used except for the 3-gram based models. This suggests that the simplified French procedure could be used in the Tel-Jeunes' system if combined with the appropriate feature type.

4 Discussion and conclusion

The first remark we could have looking at the results is that no model seems to work very well. This is not surprising as the system attempt to deal with questions related to the personal life of young people and the number of potential questions is so large that it is impossible to be sure that the proper answer to a given question is available in the bank, even if it is large. A more appropriate evaluation of the models by restricting the experiment to questions which are answered in the bank would potentially be an issue but the importance of the task makes it practically impossible at a reasonable cost as there are about 8000 Q&A pairs to look at. However, even if they seem to perform moderately, the different models could still compared and conclusions could be drawn from these results.

First, the theoretical assumption that using n-gram features would bring betters results than word features was not validated across all our experiment's results. As a matter of fact, word based models performed rather well obtaining better results, on several occasions, than models using 3-grams or 5-grams and even outclassing 4-gram based models on one measure. This result differs greatly from the ones obtained by Natrajan et al. [8] which found that word features were not suitable for retrieving garbled texts. Furthermore, in the construction phase, the QR model that obtained the highest categorization accuracy was the word based one, confirming the results obtained by Berger and Merkl [3] which found that n-gram features were not significantly better than word features for text categorization. However, in the Tel-Jeunes QR, the results do seem to indicate that using 4-gram features system would be the best choice, particularly when coupled with the simplified French procedure. Indeed, the best results on almost all performance measures McNamee and Mayfield [7] which found that 4-gram features should be used with the French language.

More importantly, our results confirmed the appropriateness of the carried out experiment as the previous choice of adopting 3-grams and combining them with the simplified French procedure for the Tel-Jeunes' system was obviously not the best one. The drawn conclusions reinforce the idea that one should be careful when using n-grams instead of words when trying to counter textual deformation problems as our experiment suggests that it is not always the right choice.

On another level, even though the elimination of infrequent term is well documented and largely accepted, its application with word based terms in contexts where textual deformations can be found raises some questions. Consider a very useful word that is from time to time erroneously written by the users. The erroneously spelled versions of this word can be considered as infrequent words and be eliminated even thought their semantic value for representing the questions is the same as the correctly spelled counterpart.

Experiments also pointed out that there seems to be a relation between the length of the question and the retrieval performance of the different models meaning that the feature type could be chosen relatively to the question's length. This is another element that could help determine which feature type or combinations of feature types to consider.

From our experiments, it seems that the simplified French procedure clearly improves the performance of all the models except the 3-gram based model. The exact reasons for this would probably be very useful to better understand the strength and weaknesses of the simplified French procedure, but also of n-grams.

References

[1] R. C. Angell, G. E. Freund, and P. Willet. Automatic spelling correction using trigram similarity measure. *Information Processing and Management*, 19:255–261, (1983).

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. Addison Wesley/ACM Press, (1999).

[3] H. Berger and D. Merkl. A Comparison of Text-Categorization Methods Applied to N-gram Frequency Statistics. *Proceedings of the 17th Australian Joint Conference on Artificial Intelligence*, (2004).

[4] R. J. D'Amore and C. P. Mah. One-time complete indexing of text: Theory and practice. *Proceedings of the 8th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1985).

[5] G. Forman. A Pitfall and Solution in Multi-Class Feature Selection for Text Classification. *Proceedings of the 21th International Conference on Machine Learning*, (2004).

[6] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge (qac-1): An evaluation of question answering tasks at the ntcir workshop 3. In Mark T. Maybury, editor, *New Directions in Question Answering*, pages 122–133. AAAI Press, (2003).

[7] P. McNamee and J. Mayfield. Character N-gram Tokenization for European Language Text Retrieval. *Information retrieval*, 7:73–97, (2003).

[8] A. Natrajan, A. Powel and J.-C. French. Using N-grams to Process Hindi Queries with Transliteration Variations. *Technical Report No CS–97–17*, (1997).

[9] G. Salton and M.J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, (1983).

[10] F. Sebastini. Machine learning in automated text categorization. *ACM Computing Survey*, 34:1–47, (2002).

[11] E.M. Voorhess. Overview of the trec 2001 question answering track. In *Proceedings of the ninth Text REtrieval Conference*, (2001).

[12] E.M. Voorhess. Overview of the trec 2004 question answering track. In *Proceedings of the 2004 Edition of the Text REtrieval Conference*, (2004).

[13] Y. Yang and L. Xin. A re-examination of text categorization methods. *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, (1999).