Extraction of Relevant Characteristics for Message Comparison

Axel Alejandro Garcia Fuentes*

Universidad Autonoma de Guadalajara
axel.garcia@edu.uag.mx

Abstract

There are aspects in the human communication that differentiate two phrases within a determined context. When the context varies the distinction between the same two phrases may change, too. Natural language is ambiguous, however, there are key elements in the natural statements that may provide enough information to fairly compare whatever two sentences in natural language. Once that capability is available for text processors computers may be capable of classify text by meaning rather than by text similarity.

I. Introduction

This document is intended to discuss the theoretical plausibility of the implementation of a system capable of extract relevant aspects of a short message in Spanish. The message length of the text to be processed should be at most 500 characters. That length was determined based on the average message length of the inquiries submitted to the Federal Institute of Information Access and Data Protection in Mexico(IFAI in Spanish)[de Acceso a la Información, 2014] . Taking that as base, it was observed that the turn around time in which the information is sent to requestors has a mode of 20 working days and 28 natural days. There are eight extra natural days requestors should tolerate when requesting information through the IFAI system. Since that is the communicated waiting period it is not really a problem but an improvement area.

The real question is are modern information technology not enough to create a system that can help trespassing the barrier of human capabilities?, is it possible to create a system that can compare a given question against a database of previously asked questions and find out what questions are close enough to mean most likely the same thing?.

The state of the art for natural language processing has achieved important progress in answering questions algorithms

II. Text Being Processed

The text being processed is written in natural language. The text is not longer than 200 characters and includes punctuation marks and the Spanish alphabet. It is written in Spanish and it is not error free which means that it may have lexical and syntactical errors. An additional characteristic the text under analysis has is it is written to request information. It was took from a government web site that offers information to the general public. The main objective of such public site is offering information access to anybody that needs it. Processing natural language has relevant challenges like identifying unknown words and determine whether that word it is a misspelled word or if it is a word that definitely does not exist for the language. There is another challenge for detecting ambiguous words and assign them the meaning that they are most likely to have with respect a context or a corpus of questions.

*A thank you or further information

III. Ambiguity

As [Monz, 2010] suggests, the words with higher inverted document frequency (idf) will be preferred to determine the topic of the short text :

$$idf(t) = log\frac{N}{df(t)} \quad (1)$$

Where N is the number of documents in the collection and *df(t)* is the number of documents in which t occurs.

Equation 1 will help to determine a list of candidate words for to identify the message topic. However, that equation measures frequency not unambiguity. Reference [de Castro Reis et al., 2012] proposes the following formula to identify unambiguous keywords:

$$\frac{df(f,c)}{\sum df(f,c')} > \alpha \quad (2)$$

> "In other words, we say that f is unambiguous if one of the concepts it may represent is alpha times more frequent in documents of the corpus than all the others combined."

That method can be used to count how many unambiguous terms a question has. That in turn can be a dimension of the description vector. That reference mentions unambiguous key terms tend to have higher IDF (Inverted Document Index) than ambiguous key terms. However, it mentions there is an overlap in the mid-lower IDF. That overlap makes it hard to use IDF to separate unambiguous from ambiguous key terms. [de Castro Reis et al., 2012] mentions IDF can be used to identify misspellings.

Coincidentally, [Monz, 2010] mentions a similar formula which is called the *presence weight* formula:

$$W_+(t) = \frac{\sum_{q' \in tsv(q) \wedge t \in q'} avg\_prec(q')}{\sum_{q' \in tsv(q)} avg\_prec(q')} \quad (3)$$

That equation compute the presence weight of the term t for a given question q. That number is proposed to be used as an index of term relevance with respect to the available corpus. That information can be used to determine if the content of any given pair of messages has the same weight with respect to the current corpus.

Both equations 3 of [de Castro Reis et al., 2012] and 2 calculate the ratio of the number of observations of t with respecto to the 100% of considered observations in the experiment. However, equation 3 is paired with the equation called *Absence Weight*:

$$W_-(t) = \frac{\sum_{q' \in tsv(q) \wedge t \notin q'} avg\_prec(q')}{\sum_{q' \in tsv(q)} avg\_prec(q')} \quad (4)$$

[Monz, 2010] suggests a statistic value called gain:

$$gain(t) = w_+(t) - w_-(t) \quad (5)$$

The gain function considers the case when the average precision of the term t is greater than the presence precision. Thus equation 5 could result in a negative number. For this work such a value means that the term t most likely is not a keyword that improve the differentiation test between two arbitrary messages under evaluation.

IV. Corpus of Questions

Each time a new question is processed it should be added to the corpus of questions since it will offer information for the next answer query. When a question is to be added to the questions and answers database, the question structure and contests are analyzed and a description vector for the question is generated. That description vector becomes the label for that question in the corpus. In that way the next question that shall be processed can be analyzed and categorized.

V. Method

[BENOTTI et al., 2014] proposes a method to interpret directions given to a system. Some parts of that method can be reused in the current research. The proposed method has two phases:

Annotation (done once)
: Associates an instruction with a system reaction. For the current investigation the description vector is the question annotation and the answer is the system reaction.

  Segmentation
  : Divide or segment the interaction into instructions and reactions. This applies in the the current work as splitting the initial text into the several questions that may be hidden into it.

  Discretize
  : Discretize reactions into canonical action sequences. Proposal is to approach discretization as the extraction of relevant characteristics from the text being analyzed. Section VI elaborates the proposed process.

Interpretation
: Predict a system reaction based on system reactions previously observed in the answers corpus.

  Filter
  : Filter the annotation-reaction pair. Keep only those pares in which the reaction can be directly executed from current state. This represent a constraint. The constraint used in this investigation is a maximum distance two questions may be to be considered equivalent.

  Group
  : Group the pairs set based on its reaction. This is group questions based on their answer closeness; i.e. clusters.

  Selection
  : Chose the group with the instructions that are the closest to the given instruction. Meaning, propose the answers of the questions with the most similar description vectors to the description vector of the question being analyzed.

The way the proposed algorithm is expected to work is by extracting a description vector from the short messages being processed and computing the distance with respect other questions previously made. As the questions and answers database grows the searches will take more time each time. To overcome that situation, the questions will be classified based on a hashing function like minhashing. The algorithm is the following:

1. Question in natural language Q is entered to the system.
2. Information extraction algorithm generates the description vector v from Q.
3. Hashing function *m()* is used to generate a hashing value h from v.
4. The hashing value h is looked for in the questions database as described in I.
5. if it is found, then the answer associated to that question is selected as proposed answer and the system will claim Q -> A
6. If it is not found, then A is empty and it is associated to h.
7. h is added to the questions database.

Whenever an empty answer is found in the system, human expert intervention is requested. Once the question is solved by the expert human, the empty entry for that answer is replaced by the provided response.

VI. Information Retrieval Dimensions

I. Content Dimensions

Ambiguity Index
: This dimension is computed with equation 5 and will allow the differentiation between two questions.

II. Shaping Dimentions

Reference [de Castro Reis et al., 2012] mentions that stop words and auxiliary language offer little information. Given that, those words will not be considered as relevant characteristics for the message comparison. Reference [de Castro Reis et al., 2012] used Yahoo! Term Extractor API, Stanford Named Entity Recognizer.

From reference [de Castro Reis et al., 2012] mentions for some real world applications they cannot afford extracting an incorrect keyword from a noisy text.

Caps First Ratio
: Reference [de Castro Reis et al., 2012] interprets Caps First Ratio as names. This can be used to differentiate a question that talks about names from another that does not.

Unambiguous Words
: Number of unambiguous worlds in phrase.

Capitalized Words
: Number of capitalized worlds in phrase(potential names).

SER
: It tells how often a given keyword shows up by itself in the search box. This is helpful because it may allow the system to detect common terms used in queries.

III. Statistics not Considered

The following metrics were proposed by [de Castro Reis et al., 2012], however, they are not considered as relevant characteristics for the underlined experiment. The reasons why that decision was made is documented in the following paragraphs.

Search Bias
: This metric relates a key word, its appearance in the question corpus and its appearance in a larger corpus. The former is referred to as logs corpus by the authors and the later as Web corpus. The text being analyzed in this investigation can be analyzed in a questions corpus. Notwithstanding that, the question corpus is the biggest scope a given term can be analyzed against. Thus there are not two levels of corpus scope to relate.

Inverse Document Frequency
: The mentioned authors suggest this metric can be used for ambiguous terms detection. For the present investigation 5 is being used with that purpose.

Session Inverse Document Frequency
: SIDF works over a user session. It consideres the query stream of such session as a document. The text being analyzed in this investigation is not meant to change nor gotten from a live stream.

VII. Test Set Definition

The test data has been obtained from [de Acceso a la Información, 2014].

References

[BENOTTI et al., 2014] BENOTTI, L., LAU, T., and VILLALBA, M. (2014). Interpreting natural language instructions using language, vision, and behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Multiple Modalities in Interactive Systems and Robots*, Volume 4 Issue 3(13):22.

[de Acceso a la Información, 2014] de Acceso a la Información, I. F. (2014). Sistema infomex gobierno federal.

[de Castro Reis et al., 2012] de Castro Reis, D., Goldstein, F., and Quintao, F. (2012). Extracting unambiguous keywords from microposts using web and query logs data. In *Making Sense of Microposts (#MSM2012)*, pages 18–25.

[Monz, 2010] Monz, C. (2010). Machine learning for query formulation in question answering. *Natural Language Engineering*, 17(4):425–454.