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. Method

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.

*A thank you or further information

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.

III. Information Retrieval Dimensions

The function of the words the processed text is composed of is different. Some of those words hold more information than others. As [Monz, 2010] suggests, the words with higher inverted document frequency (idf) will be preferred to determine the topic of the short text :

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

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:

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

> "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:

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

That equation compute the presence of weight of the term t in a given question q. It uses the frequency of the term t in the corpus of questions. Each variant of the question q is represented by q' and the function *tsv(q)* maps all the variants of q. This is a model of the average precision of the query variants in which the term occurs.

Equation 3 of [de Castro Reis et al., 2012] is similar to 2 in that both are calculating the ratio of the number of observations of t in the 100% of observations considered for the experiment. This will be used as a proposed dimension called *"Presence Weight"* since the authors of equation 3 used that name.

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. This can be another dimension.

Reference [de Castro Reis et al., 2012] shows that stop words and auxiliary language is less often in query streams. If those elements are being used less then they offer little information and can be neglected. The current work proposes to ignore stop words during description vector creation.

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.

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.

References

[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.