# A comprehensive information extraction module for reducing call handling time in a contact centre

**K.I.H. Gunathunga, Y.H.P.P. Priyadarshana, K.K.A. Nipuni N. Perera, L. Ranathunga, P.M. Karunaratne, T.M. Thanthriwatta**

Faculty of Information Technology, University of Moratuwa, Sri Lanka

***Abstract-*** Information extraction plays an important role in text related research and application areas such as text mining and dialogue systems. Information extraction can be done using key word extraction and measuring the semantic similarity between texts. These concepts are applied to address a key issue in the telecommunication contact centre domain where the customer dissatisfaction is increasing due to higher call handling time. The proposed method is a combined with a key word based approach and a semantic similarity based approach with the use of semantic nets. The semantic similarity of two sentences is calculated using word similarity and the word order. Experiments on two sets of sentence pairs illustrates that proposed method provides a similar measure which is significantly correlated to human intuition. The overall accuracy of the information extraction module is approximately 70% based on the evaluation results.

***Index Terms-*** Information extraction, natural language processing, semantic nets, sentence similarity

## I. INTRODUCTION

Telecommunication is a highly competitive and evolving industry. Organizations in this industry handle large amount of data generated through different operations which plays a key role in enterprise decision making. Authors have identified a key problem in telecommunication contact centre domain in Sri Lankan context which can be addressed through an application of natural language processing. The key objective of the proposed solution is to reduce the call handling time in the contact centre and thereby increase the customer satisfaction and minimize customer churn.

After having a detailed discussion with contact centre agents, authors identified the limitations and the drawbacks in the existing system. In a contact centre, call routing among contact centre agents is an important task. Since the contact centre handles thousands of inbound calls daily, the call routing mechanism should be more efficient, especially in peak hours. International standard average call holding time is 3.5 minutes, but it is set to 5 minutes in Sri Lankan context with the capacity of available human resources. Customers have to explain their problems within 5 minutes and once the call duration exceeds the time limit the call get terminated automatically. If a customer has not completed the conversation then he has to call again to the contact centre. Most of the time these repeated calls are not routed to the same agent who carries out the previous conversation with the customer. In such case, customers have to explain their problem again to a new agent. This may create a bottleneck in contact centre queues, which has become a major reason for the customer churn. According to recent findings customer rate of shifting from PSTN (Public Switch Telephone Network) lines to mobile service providers has increased. Therefore it is essential to enhance the existing system with reducing the drawbacks on it.

The proposed solution is based on an information extraction mechanism in order to provide a brief understanding to contact centre agents, regarding earlier conversations happened between customers and fellow agents. The initial conversation between a customer and a contact centre agent is converted into a text by using CMU Sphinx speech recognition toolkit[1]. The generated text file is provided as the input for the information extraction module. Extraction contains important facts in the conversation such as name of the agent who handles the call, type of the problem and a set of actions performed by both customer and the agent which relates to the problem. This paper is based on extracting set of actions from the text with the usage of natural language processing techniques.

## II. METHODOLOGY

The proposed method is a combination of key word based approach and a semantic similarity based approach which obtains information with the assistance of semantic knowledge base Figure 1 shows the procedure for extracting information by combining two approaches.

### *A. Key word based approach*

In this approach, different sets of key words are defined based on their relatedness to the service categories provided by the contact centre as follows:

- Broadband/ADSL category
- PSTN category
- CDMA category
- PeoTV category

Key words are categorized with the knowledge gain by interviewing the contact centre agents. For each service category, agents have a pre-defined set of questions. When a customer connects with an agent, these questions have been asked and based on the given answers of agent has to identify the exact problem of the customer. By considering the sequence order of the questions and the actions associated with them, unique tree structures are defined for each service category.

[1] http://cmusphinx.sourceforge.net/

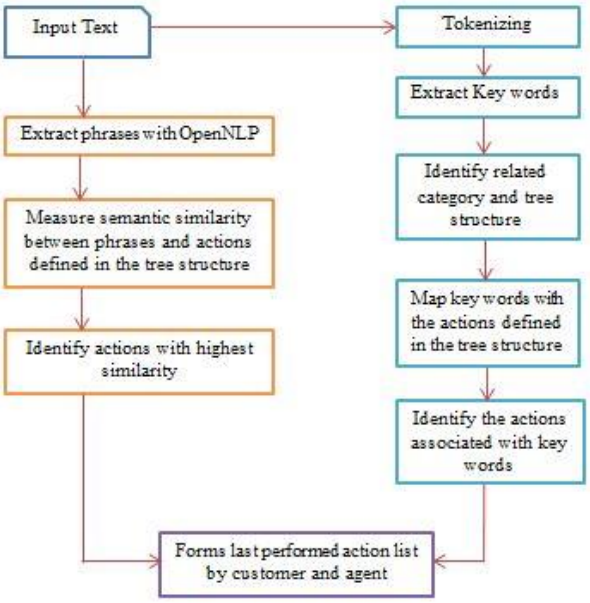


Figure 1: Information extraction module

In the first step, the text file is tokenized. Authors have used Apache OpenNLP API for the tokenizing process. All tokens are compared with pre-defined key word sets and forms a joint word set which contains common and distinct key words, for each category. Let's take $T$ as the token set in the text file and $S_i$ as the pre-defined key word set associated with each category. Joint key word set $J$, is defined as $J = \{T \cap Si\}$

Four different joint sets are formed and the one with most number of elements is selected. Related service category and the selected key word set are considered for further processing. After identifying the key words in the text file, they are mapped with the relevant tree structure which is created for service category. All the nodes which are associated with the extracted key words are identified by traversing through the tree. At the end of this process a set of actions described in the text file can be extracted.

```
broadband/adsl problem
    internet connection is disconnected
        noise in telephone line
            report as pstn problem
                issue a reference number
                    inform maintanance staff
        disturbances in telephone line
            report as pstn problem
                issue a reference number
                    inform maintanance staff
    internet connection is not disconnected
        internet connection is slow
            exceed gb limit
                add more gb
                transfer package
            not exceed gb limit
                issue a referance number
                    inform maintanance
        ask router type
            zt/prolink/dlink/tplink
                ask bulb status
                    internet bulb status
```

Figure 2: Portion of the tree structure created for broadband service category

Key word based approach is not sufficient to extract the exact set of actions in a text file. Since the speech recognition process not providing accurate results, there may be a possibility of missing some key words in the text file. This emphasizes the important to introducing another approach which increase the effectiveness of the information extraction process. Therefore an approach based semantic similarity between short texts and sentences is proposed.

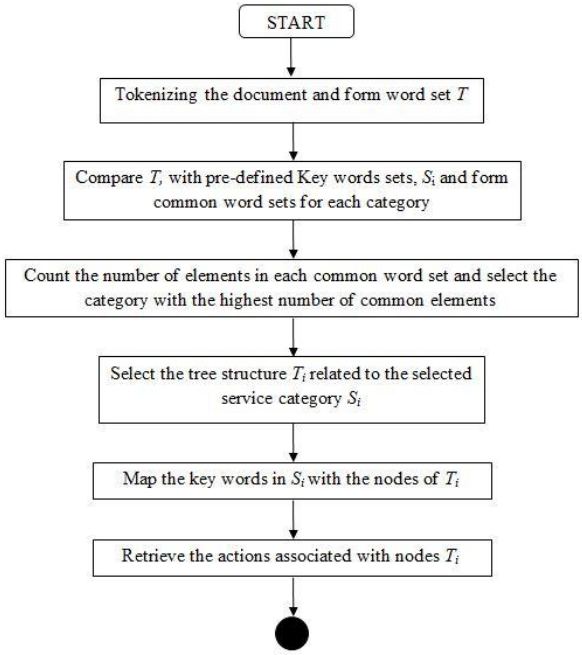


Figure 3: Flow chart for key words based approach

*B. Semantic similarity based approach*

In this approach, the semantic similarity between the extracted text phrases from a text file and node values in a related tree structure is measured. Apache OpenNLP [5] model trainer is used to extract key phrase from a text file. Authors have prepared a contact centre domain specific training data set by conducting discussions and interviews with contact centre agents, listening to recorded conversations and analyzing the Frequently Asked Questions (FAQ) in telecommunication domain. Training data set contains the possible ways of explaining customer problems and answers provided by contact centre agents.

The proposed method measures similarity between two sentences based on semantic and syntactic information includes in compared texts. A sentence or a text phrase is considered as a sequence of words. Each word contains useful information along with their combination which makes a specific meaning.

Figure 4 presents the procedure for computing the similarity between two sentences. Instead of using a fixed set of vocabulary like in existing methods, the proposed method forms a joint word set which contains all the distinct words in two sentences. For each sentence, a semantic vector is formed with the assistance of a semantic database. Semantic similarity is calculated using two semantic vectors.

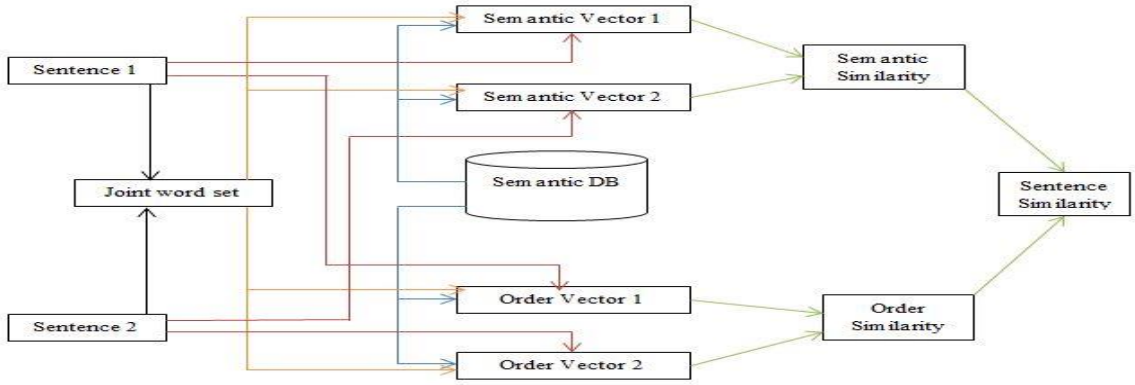


Figure 4: Sentence similarity based approach

A word order vector is formed for each sentence by considering the sequence of words in sentences. The word order similarity is calculated using two word order vectors. Finally, the combination of semantic similarity and word order similarity is used to compute the sentence similarity. The following sections provide detailed descriptions of each step in the process.

*C. Measuring semantic similarity between words*

The proposed method is based on a hierarchical semantic knowledge base which is important to determine semantic distance between words. Available knowledge bases consist of a hierarchical structure and models human common sense knowledge for different domains. Let's consider two words $W_1$, $W_2$ and their semantic similarity $S(W_1, W_2)$. Authors have used WordNet, an available lexical database where the words are organized into synonym sets (synsets) in the knowledge base, with semantics and relation pointers to other synsets [1]. One direct method to measure the similarity is finding length of the shortest path connecting two words. But this method may provide less accurate results when it applies to ore general semantic nets such as WordNet [1]. To overcome this drawback, the direct path method has to be modified. Authors have used the method which was determined by the method proposed in [2]. It is clear that words at upper levels in the hierarchy has more general semantics and less similarity among them, whereas the words at lower levels have more specific semantics with more similarity. Therefore, the depth of words in the hierarchy is considered in measuring semantic similarity. In summary, Semantic similarity between words $W_1$ and $W_2$ can be defined as transfer function of path length and depth.

$$S(W1, W2) = f(l).f(h)$$

where,
$l$ = shortest path length between $W_1$ and $W_2$
$h$ = depth between $W_1$ and $W_2$

When considering the transfer functions, the similarity is varying from *exactly the same* to no *similarity* [2]. If we assign 1 to the exact similarity and 0 to no similarity then the interval of similarity is [0, 1]. When the path length is decreasing to zero, the similarity is increasing towards limit 1. And when path length is increasing infinitely, the similarity is decreasing towards 0.

This behavior emphasizes that the transfer function must be a nonlinear function. Based on these considerations $f(l)$ is defined as decreasing function of $l$ as follows;

$$f(l) = e^{\alpha l}$$

Where α is a constant. The purpose of representing the function in exponential form is to satisfy the constraint of keeping the value of $f(l)$ within the range of 0 to 1. For WordNet the proposed value for α = 0.2 as reported in [3].

In the same way $f(h)$ can be defined by considering the behavior of words at upper levels of hierarchical semantic nets are more general and have less similarity between words than lower levels. As a result, $f(h)$ is defined as an increasing function of $h$.

$$f(h) = \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}$$

where β > 0 is a smoothing factor as β → ∞, then the depth of a word in the semantic net is not considered. For WordNet the proposed value for β = 0.45 as reported in [3]. The optimal values for α and β depend on the knowledge base used and are determine with human similarity ratings [2].

*D. Semantic similarity between sentences*

Sentence is a collection of words and it is reasonable to represent a sentence using its words. In proposed solution, a semantic vector is formed dynamically as suggested in [2]. A joint word set $T$ is formed for given two sentences $T_1$ and $T_2$.

$$T = T1 \cup T2$$

The joint word set $T$ {$W_1$, $W2$… $Wm$}, contains all distinct words from $T_1$ and $T_2$. Each sentence is represented by using the joint words set. The semantic vector $S_i(i=1,2,\ldots,m)$ is formed based on the semantic similarity of the corresponding words to a word in the sentence. Let's take $T_1$ as an example.

Case 1: If $W_i$ is contained in $T_1$, $S_i$ is set to 1
Case 2: If $W_i$ is not contained in $T_1$, a semantic similarity measure is computed between $W_i$ and each word in the sentence $T_1$. The most similar word in $T_1$ to $W_i$ is the one with highest

similarity score θ. If θ exceeds a preset threshold, then $S_i$ = θ; otherwise $S_i$ = 0.

Since the word similarity is measured between different words, the highest similarity score may be very low. This may indicate as that words are highly dissimilar. To avoid this dissimilarity in the semantic vector, a threshold is used. For WordNet this preset threshold is set to 0.2 [2]. As suggested in [2], information content of words has considered for increasing the accuracy of similarity measure by using the brown corpus. Since the brown corpus is outdated, it is not applied to the method proposed in this paper. Finally, the semantic similarity between two sentences is represented as the cosine coefficient between two semantic vectors.

$$Ss = \frac{S1.S2}{\| S1 \|.\| S2 \|}$$

*E. Word order similarity between sentences*

If two sentences contain same words, any method based on "bag of word concept" will decide that both are exactly same [2]. But through a human interpretation it can be showed that two sentences are similar only to certain extend. This is stated the importance of considering the word order of sentences in measuring the similarity.

Let's consider two sentences $T_1$ and $T_2$ and its joint word set as $T$. Each word in $T_1$ is assigned a unique index number. This indexing is simply based on the order of appearance of each word in the sentence. A word order vector $R$ is formed for each sentence based on the joint word set. For each word in $W_i$ in $T$, the most similar word in $T_1$ is considered.

Case 1: If $W_i$ is appeared in $T_1$, then fill the entry for this word in $R_1$ with the corresponding index number of $T_1$. Otherwise find the most similar word $W_r$ in $T_1$.

Case 2: If the similarity between $W_i$ and $W_r$ is greater than a preset threshold, then the entry of $W_i$ in $R_1$ is filled with the index number of $W_r$ in $T_1$. If not the entry of $W_r$ in $R_1$ is 0.

By following this procedure two word order vectors, $R_1$ and $R_2$ are formed for $T_1$ and $T_2$.

Authors have followed a measure for measuring the words order similarity between two sentences as suggested in [2].

$$Sr = 1 - \frac{\| R1 - R2 \|}{\| R1 + R2 \|}$$

*F. computing the overall sentence similarity*

Overall sentence similarity is a combination of semantic similarity and the word order similarity. But when combining two results the relative contribution of them has to be considered. Therefore the overall similarity is defined as:

$$S(T1, T2) = \mu Ss + (1 - \mu)Sr$$

where μ < 1 denotes the relative contribution. According to the experimental results mentioned in [2] word order threshold is set to 0.4 and μ is set to 0.85. Authors have enhanced the method proposed in [2] by considering most similar pair of synsets for each word, instead of picking the first noun synsets. The main reason for this enhancement is a word can be mapped to multiple synsets and finding most similar pair denotes the human tendency for pattern seeking. Algorithm for the enhancement is as follows:

```
function get_best_synset_pair(w1, w2)
w1 = first word
w2 = second word
max_sim = maximum similarity score between two synsets
best_pair = most similar synset pair
synsets1 = synsets for w1
synsets2 = synsets for w2

begin
best_pair ← none
max_sim ← -1.0
if (length of synsets1) == 0 OR (length of synsets2) == 0
        return None
else
foreach syn1 ϵ synsets1
foreach syn2 ϵ synsets2
                sim = path_similarity(syn1, syn2)
if sim > max_sim
max_sim ← sim
best_pair ← syn1, syn1
end foreach
end foreach
return best_pair
end
```

Figure 5: Pseudo code of the algorithm to get the best synset pair

## III. RESULTS

To implement this solution, authors have used python and Natural Language Toolkit (NLTK). Since sentence similarity is highly depends on semantic similarity between words in the sentence, it is important to measure the accuracy of the word similarity algorithm. Authors have compare the method proposed in [2] with newly proposed method.

Let's rename method proposed in [2] as $ALGO_{old}$ and new method as $ALGO_{new}$. Following a similar procedure to Miller and Charles [4], a subset of 15 word pairs are considered for the comparison between two algorithms.

Table I: Word similarity results comparison

| Word pair | Similarity score for $ALGO_{old}$ | Similarity score for $ALGO_{new}$ |
|---|---|---|
| [autograph, shore] | 0.29 | 0.16 |
| [autograph, signature] | 0.55 | 0.82 |
| [boy. lad] | 0.66 | 0.82 |
| [boy, sage] | 0.51 | 0.37 |
| [cock, rooster] | 1.0 | 1.0 |

| | | |
|---|---|---|
| [cord, smile] | 0.33 | 0.13 |
| [cord, string] | 0.68 | 0.82 |
| [forest, woodland] | 0.70 | 0.98 |
| [forest, graveyard] | 0.55 | 0.20 |
| [hill, woodland] | 0.59 | 0.36 |
| [hill, mound] | 0.74 | 0.99 |
| [implement, tool] | 0.75 | 0.82 |
| [midday, noon] | 1.0 | 1.0 |
| [magician, oracle] | 0.44 | 0.30 |
| [magician, wizard] | 0.65 | 1.0 |

The accuracy of the experimental result can be elaborated as follows. According to the Table I the computed similarities using proposed new method is lined up with intuition. For example, similarity between words *autograph* and *signature* is higher than the similarity between *autograph* and *shore*. The similarity between words *magician* and *wizard* is higher than *magician* and *oracle*. This clearly shows that the proposed method is providing better results, when comparing to the existing methods. Therefore, the accuracy of the word similarity measurement is caused to increase the accuracy of semantic similarity based approach to extract information.

Accuracy of the Key word based approach is measured based on the comparison between the proposed solution and human involvement method where authors listen to the audio clips that received from contact center. 50 audio files (a) have used for the experiment. For each audio file the total number of actions mentioned in the audio file and the number of correctly identified actions through the proposed solution is considered. These two values are considered to measure the precision of the proposed method.

$$precision\ (a) = \frac{Ca}{Ta}$$

Where,
$Ca$ = Number of correctly identified action in an audio file
$Ta$ = Total number of actions in an audio file

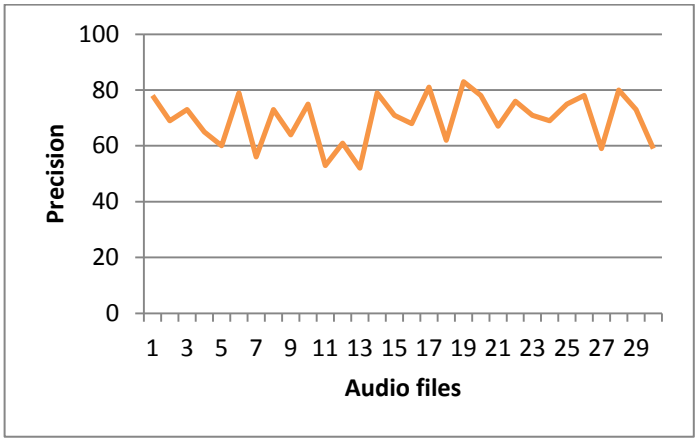


Figure 6: Precision trend

## IV. CONCLUSION

When evaluating the experimental results on information extraction module, accuracy of getting the correctly identified actions is approximately 70%. Authors have compared this approach with the actions identified with a human involvement by listening to audio files. But in the actual contact center domain the gents do not have the access to listen previously made calls by a particular customer to the contact center, in order to complain regarding their problems. In such case agents who handle repeated calls do not have a prior knowledge regarding customer's previous experience. But through the proposed solution it provides at least a basic understanding about the customer's problem type and a set of actions taken to overcome the problem. This provides benefits for both contact centre and its customers in different aspects. From the contact centre point of view, it reduces the call handling time, minimize the number of repeated calls made by customers for the same complain and increase the number of complains that a contact centre can efficiently handles within a day. At the same time it makes customers life easier by supporting agents to solve complains quickly and reducing the waiting time in contact centre queues. Ultimately it increases the overall customer satisfaction and reduces the customer churn which is a key objective of the proposed solution.

In this solution, WordNet as the hierarchical semantic knowledge base, with general English usage, in order to measure the semantic similarity between sentences. As future enhancements, authors will work on build a domain specific hierarchy for the call center domain and combined it with existing semantic nets, which gives more precise and accurate measurements.

## ACKNOWLEDGMENT

The authors want to acknowledge the support and collaboration received from the administrative staff of Sri Lanka Telecom PLC and both academic staff and students of Faculty of Information Technology, University of Moratuwa, Sri Lanka.

## REFERENCES

[1] D. Yang and D.M.W. Powers, "Measuring semantic similarity in the taxonomy of WordNet," In Proc. of the 28th Australasian Comp. Sci. Conf., pp.315–322, 2005

[2] Li, D. McLean, Z.A. Bandar, J.D. O'Shea, and K. Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics", IEEE Transactions on knowledge and data engineering, Vol. 18, August 2006.

[3] Y.H. Li, Z. Bandar, and D. McLean, "An Approach for Measuring Semantic Similarity Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/ Aug. 2003.

[4] G.A. Miller and W.G. Charles, "Contextual Correlates of Semantic Similarity," Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28,1991

[5] G.S. Ingersoll, T.S. Morton, and A.L. Farris, "Taming Text: How to Find, Organize, and Manipulate It"

## AUTHORS

**First Author** – K.I.H. Gunathunga, Undergraduate of Faculty of Information Technology, University of Moratuwa, Sri Lanka, isuruhasarel@gmail.com
**Second Author** – Y.H.P.P. Priyadarshana , Undergraduate of Faculty of Information Technology, University of Moratuwa, Sri Lanka, toprasanyapa@gmail.com
**Third Author –** K.K.A. Nipuni N. Perera,Undergraduate of Faculty of Information Technology, University of Moratuwa, Sri Lanka, nipuninamali@gmail.com
**Fourth Author** – Dr. L. Ranathunga, B.Sc. Sp(Hons), M.Sc., PGDip in DEd. (IGNOU), PhD (Malaya), MIPSL, MCSSL, Senior Lecturer, Head, Department of Information Technology, Faculty of Information Technology, University of Moratuwa, Sri Lanka, lochandaka@uom.lk
**Fifth Author** – P.M. Karunaratne, MBA, Msc, B.Sc.Eng. , Senior Lecturer, Head, Department of Interdisciplinary Studies, Faculty of Information Technology, University of Moratuwa, Sri Lanka, pmkaru@itfac.mrt.ac.lk
**Sixth Author –** T.M. Thanthriwatta, B.Sc.(Hons) in IT, Lecturer, Department of Information Technology, Faculty of Information Technology, University of Moratuwa, Sri Lanka, thilinat@uom.lk

**Correspondence Author** - K.I.H. Gunathunga, isuruhasarel@gmail.com, +94779283819