J. Basic. Appl. Sci. Res., 2(8)7572-7582, 2012

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

An Intelligent Agent Based Model for Cross-Language Retrieval of Chunks using Linguistic Corpora

***Amin Nezarat[1], Tayebe Mosavi Miangah[2]**

[1] Department of Computer, Yazd Branch, Islamic Azad University, Yazd, Iran
[2] English Language Department, Payame Noor University of Yazd Branch, Iran

ABSTRACT

Cross-language information retrieval is a retrieval process that the user can present queries in one language to retrieve documents in another language. As users often are not able to select and find good equivalents for their target literature, then it is profitable to make use of some linguistic resources such as corpora in addition to bilingual dictionaries. The problem is still more easily to be solved in a case that the subject of each query is to be determined and dealt with individually. For this purpose, a classification method has been introduced in which classes are identified first, and then new documents are automatically placed in related topics after being processed according to the new retrieved documents. Moreover, a bilingual lexicon of English and Persian parallel chunks from two large monolingual and bilingual corpora have been constructed which can be directly applied to cross-language information retrieval tasks.
The results gained from an experiment which was performed on a set of one hundred English and Persian phrases and collocations demonstrated that our system is highly effective in assisting the users to find the most relevant and suitable equivalents of their queries in either language.
KEYWORDS*:* chunk retrieval, cross-language information retrieval, linguistic corpora, text classification, Persian language, Intelligent agent

I. INTRODUCTION

Information retrieval (IR) is a crucial area of natural language processing (NLP) and can be defined as finding the documents whose content is relevant to the query need of a user. Cross-language information retrieval (CLIR) refers to a kind of information retrieval in which the language of the query and that of searched document are different. In fact, it is a retrieval process where the user presents queries in one language to retrieve documents in another language[9].

In this paper we construct a bilingual lexicon of English and Persian parallel chunks from two large monolingual and bilingual corpora which can be directly applied to cross-language information retrieval tasks. For this purpose we use a statistical measure known as Association Score (AS) to compute the association value between every two corresponding chunks in the corpus using a couple of complicated algorithms. Once the CLIR system was developed using this bilingual lexicon, we will perform an experiment on a set of one hundred English and Persian phrases and collocations to see to what range our system is effective in assisting the users to find the most relevant and suitable equivalents of their queries in either language.

II. THE PROBLEM OVERVIEW

Impossibility of data retrieval is one of the main problems of Persian users in cyber space. This problem is largely caused by small volume of Persian data in this space. Using of dictionary cannot be very helpful because it cannot make a good deal with chunks (compound words). By examining the Persian keywords in Google ,we notice that users often could not select and translate the proper equivalents for their topics. In fact, the main problem is the users' unawareness of these chunks proper equivalents (combinations of words) in Persian. For example when we search the Persian term" به خدمت ارتش در آوردن" in Google Translate, the term of "Army service in bringing" is presented as its equivalent .This term (which is not its correct equivalent) is translated using the dictionary and without considering the user's context. The number of Persian words combinations is very much and few dictionaries can help a translator in this regard.

Hence, creating the database and the corpus which contains more accurate Persian to English translation of chunks is necessary in cross-language data retrieval engines. Besides the introducing a new method in text clustering and making the equivalents of translational terms ,we can use a database including all Persian chunks and its English equivalent with cross-language search software as a complete and comprehensive database in translation engines .

***Corresponding Author:** Amin Nezarat, Department of Computer, Yazd Branch, Islamic Azad University, Yazd, Iran. No B12,Almas Building,Khoram St.,Yazd,I.R.Iran.00983517252684amin_nezarat@hotmail.com

III. RESEARCH QUESTIONS

As mentioned above, the goal of this study is implementation of a Persian to English cross-language retrieval system which can solve the translational ambiguities in chunks. Hence, this study tries to answer these following questions:

- Can cross-language retrieval system be implemented just with the help of dictionary?
- Can intelligent data collecting agents be used for making a binary language corpus?
- To what extent is the role of corpus in cross-language data retrieval?

IV. RESEARCH METHODOLOGY

In this research, which is proceeded with the goal of getting one Persian to English binary language corpus, we first collected different words and texts from numerous websites by using an intelligent kind of software called "Agent/عامل".Then, the collected data fell into the different groups by using the clustering methods. The resultant groups are topically differentiated and have various features which are usable in translation software, too. The present study is of implementation and development type. In addition to providing a laboratorial software in this study, the genuine specimens calculated as parameters and authenticity of algorithm and presented model have been examined .The basis of this study is experimental-analytical which is based on implementation of laboratorial model.

V. INFORMATION RETRIEVAL

Information Retrieval indicates to a complex technology of searching and extracting information, data and meta-data in various kinds of informational resources such as databases, set of pictures and the web [3].

Increasing the body of information stored in different resources, the process of information retrieval and the extraction methods have special importance.

This information may include any format like text, image, voice and video. Unlike databases, information stored in large information sources such as web and its subsets don't comply with determined structure like social networks and generally don't have defined and intended meanings. The objective of information retrieval in such conditions is to help user find desirable information from a set of unstructured information [1,10].

Information retrieval is also used to organize and restructure information retrieved. It includes thematic classification of information obtained in terms of text or document contents.

In this paper a classification method is to be introduced in which classes are identified first, and then new documents are automatically placed in related topics after processing according to new retrieved documents.

In this approach, the system identifies subject of each text and use different methods to insert that document in a proper group or class which, in turn, leads to increased information retrieval speed searching among million documents available in computerized systems.

A. Classification method

In the field of text mining (TM), classification techniques are used in selecting documents. In addition, these techniques are used for visibility of results. It is, for instance, very helpful to use these techniques when the presentation of cases in separate groups simplifies interpretation of results.

These "classification" techniques extensively studied in "information retrieval" can be divided into two general approaches: bordering approach and hierarchy approach. In "bordering classification", many classes are specified and these classes remain unchanged during the classification process. In the first step, classes are selected in a random way and developed during a repetitive process[4].

In "hierarchy classification", classes are repeated, changed and adjusted during this process.

The first classification algorithm was "K-Means" algorithm, the main principles of which is on this basis that each "class" is represented by "mean" of all its members and this can be a basis of reclassification process and can be repeated frequently. In addition, we can use other statistical presentations like "median" or "mode" instead of "mean". Other algorithms have been developed from family of "K-K-Means" algorithm (as a sample of algorithm). Executing such algorithms can be observed in CLARA or CLARANS systems[5].

"Text classification" is a process with the aim of assigning pre-defined classes to documents. This classification process can be formulated with assigning Boolean values to any couple arranged as form [class, document] which is performed in terms of some selected standards.

There are many for classification. Among all approaches the emphasis is on the traditional statistical classifiers such as the nearest neighborhood algorithm (KNN) and NB algorithm because they have the best amount of "complexity/performance" coefficient. They represent good performance, keep their simplicity and execute in an efficient form[5,2].

There are various methods for measuring the relationship between two documents or texts which can be used in classification techniques.

Standard approaches that evaluate the accuracy of response in retrieval systems of information for structured and unstructured texts. A retrieved text is considered to be related when is in accordance with user's information requirement rather than containing only searched words. For example, when user searches word "python", she/he may need information about python or may require information about ionic program language of python. It is hardly possible for a system to find a suitable response by merely one word. We can use methods to measure relationship between user's requirement and retrieved texts [1].

VI. INTELLIGENT AGENTS

An agent can be a person, machine, a part of software code or anything else and its dictionary definition is expressed as :Anything that has the ability of doing something. Agent is a software system with features such as "understanding and change in its text", autonomy, adaptation(response to the text changes, learning) and sociality .An agent has "previous information of text"," previous experiments which can help learning" , "a goal which it should try to achieve" and "information and observation of its surroundings" at the same time proceeds an action .In addition to the above mentioned features ,artificial intelligence researchers believe that agent is a computer system with humane features such as knowledge, belief, determination and commitment. In this view ,agent has following additional features:

-Displacement: The agent can displace throughout an electronic network.

-Sincerity: The agent doesn't transfer incorrect information intentionally.

-Benevolence: The agents objectives aren't in conflict with each other and each agent tries to do only its own task.

-Rationality: An agent attempts to fulfil its own purpose[8].

VII. THE STRUCTURE OF AN OUR INTELLIGENT AGENT

The agent contains program and architecture. Program is a function which implements agents behaviour. In other words, it deals with operation implementing from agents understanding to a certain behaviour. The artificial intelligence task is designing operating program. Computational hardware which operating program implemented are called "operating architectures". This architecture can be a simple computer or it can contain special software equipments such as audio and visual processing systems. Hardware should also have a software which is located between computer and operating program and make the high level programming possible. The architecture presents its observations to the agent by sensors, runs the program and transfer agents behaviour to the environment and other agents through executive body [6].

In order to complete necessary database in data retrieval software, an intelligent software agent is provided which can collect the data and text and at the same time displace in web environment. The initial structure of this intelligent agent is discussed in the next chapter.

In order to determine starting point of software agent, Hamshahri (www.hamshahri.com) newspaper site was considered as input origin. since type of texts should be determined in structure designed for Persian corpus ,the agent movement was variably done (politics, religion, thought and etc). After reading each text in web pages related to the designated section , the text entered in corpus and the type of sentence was also defined. since each news was associated with some other related news, the related web page entered next link and its writing content added to the corpus and continued to the complete reading of the whole site. This was done by reading HTML and XML structures. After finishing the task of this part and experimental observation of retrieved texts, software bugs were identified and fixed. It is worth mentioning that this agent acts as a robot and after releasing in web start collecting necessary required data and send retrieved documents to the textual corpus. In continuation, a list of valid sites with formal writing style was selected and presented to software agent as "driving list" in order to increase volume of textual corpus. Then movement command issued using GO command and a software started writing ,recording and moving to the next web page and website according to a predefined program. Considering proceeded programming in software agent, this task was completely done without hand interference and automatically. Different types of classified texts in Persian textual corpus are located in following groups:

- Politics-medicine-literature-sport-art-religion-science-events- social and economics,…

We confront high volume of unrelated data during reading web pages such as pictures, charts, and links to other pages and delete them after determining their types. thus ,we first change read structures to the standard XML and tag formats and after inserting in textual corpus we see that main used tags in textual XML structure are as table 1:

Table 1 : list of corpus fields

ID	**شماره جمله**
Text	متن
Type	نوع متن
Link	آدرس صفحه
Date Time	زمان و ساعت

Considering high volume of collect data, SQL Server 2005 software agent database was selected in order to increase the speed of data retrieval in corpus. This sever was selected because of its high ability in processing questions and SQL Query.

In addition to this database ability of adaptation with lots of programming languages such as C# , the structure of saving database on hard disks as clustered forms is another reason that it has been selected.

VIII. DEVELOPING DATABASE

are Most of text information retrieval systems are relational databases. In these systems the queries are often retrieved from unstructured contexts called non-characteristic data. In other word, databases are designed to generate relations between data and a set of records have predefined characteristics.

The main difference between the two systems of retrieval model and information collection includes related query language and data structure. The main resources for gathering structured data contain digital libraries and specialized web logs which are classified and distinct.

In this research we focus on such resources to complete our two linguistic corpora. One example of these sources is indexing base for students' articles and theses. These indexing bases mainly bilingual.

IX. EVALUATION OF RETRIEVED DATA

There are various methods for evaluating the relation of two documents or texts which can be used in clustering discussion.

The standard approach in evaluating the authenticity of answer in data retrieval systems is the measurement of texts relation or lack of relation..[3].A retrieved text is related when it can meet users informational needs .For example ,when a user search the word" python"

Is it clear that he is looking for some information about a kind of snake or is looking for information about ionic program of python?

It is hardly possible for a system to find a suitable answer using just a word. In order to increase the rate of authenticity of answers ,we can use methods of measuring the rate of relation between users demand and retrieved texts. Thus, we can use binary decision making method [5].

Using test collections is one of the common method for measuring the rate of relation between outputs and is measurable by testing systems.

Some of the test collections which are usable in data retrieval systems are introduced in this part. We can use these collections in clustering texts. We call these collections (corpus) in linguistic discussions which contains high volume of clustered texts.

- Cranfield collections: This collection is one of the oldest measuring collections made in England in 1950. It contains 1398 abstracts from airplane magazines articles and 225 queries in this regard. Today it is barely used because the volumes of available data are too low to rely on.

- Text Retrieval Conf: Standard and technology association of America (NIST) started collecting and making a big collection in this regard. A large extend of texts collected in this collection which is more than 1/89 millions texts clustered in 450 topics .This corpus is one of the biggest available corpuses which can be used in researches.

- NII Test Collections for IR systems (NTCIR): This corpus was made during a project which was being proceeded for NIST organization of USA .In this corpus which has been developed for supporting languages of countries located in the east of Asia, the great emphasis is on cross-language data retrieval .The queries have been designed on the basis of collected texts from one language in front of other languages.

-Cross-Language Evaluation Forum (CLEF):

This corpus is also provided in European languages to help cross-language data retrieval.

-Reuters-21578:This collection contains 21578 articles from this newspaper which contains 806791 texts. This corpus is very beneficial and used a lot.

Evaluation of disordered retrieved collections:

Precision and Recall are two basic and common methods in measuring efficiency rate of data retrieval systems. We first examine these methods with a simple mode in data retrieval.

Precision (P):The percentage of related retrieved texts

$$\text{Precision} = \frac{\#(\text{relevant item retrieved})}{\#(\text{retrieved item})} = P(\text{relevant retrieved})$$

$$\text{Recall(R)} = \frac{\#(\text{relevant item retrieved})}{\#(\text{retrieved item})} = P(\text{relevant retrieved})$$

These parameters are presented in a relational state as followed:

	Relevant	**Non relevant**
Retrieved	True positive (tp)	False Positive(fp)
Not retrieved	False negative (fn)	True Negative(tn)

P=tp/(tp+fp)

R=tp/(tp+fn)

X. OUR CLASSIFICATION METHOD

Until now, we became familiar with concepts related to information retrieval and the relevancy of a document with an input query. However, another important subject here is text classification in similar groups by the help of which we can identify the relationships between every two translated sentences and find the accuracy of translation in both Persian and English sentences.

In order to determine the relationship between texts, we can consider a set in which X is a set of texts and a fixed set of classes with characteristic $d \in X$ Such classes or clusters have been defined by human beings. We also consider a set of instruction with headline D. Texts are labeled and their classes are specified, so, we can show them as (d, c).

For example,

$C = \{c_1, c_2, ..., c_j\}$

$(d, c) \in X \times C$

(d,c) = {Professor Farshchian's art works are shown in Tehran museum , Art}

In this learning method, text classification is carried out by a human agent referred to as "supervisory learning". In this learning method, a human agent classifies and labels texts. However, we can use statistical methods related to classification in order to measure accuracy of classification performed.

The method selected and adopted in this paper is X^2 method. In statistics, X^2 value represents the rate of independence in two occurrences. The two occurrences A and B are independent if and only if there are equations P(AB)= P(A) P(B) or P(A B)= P(B), P(A B)= P(A), we can obtain X^2 as follows through the following formula:

$$X^2(D,t,c) = \sum_{e_t \in \{\varnothing,1\}} \sum_{e_c \in \{\varnothing,1\}} \frac{(N e_t e_c - E e_t e_c)^2}{E e_t e_c}$$

which N value is in fact the number of all sentences observed in D, and E is the expected numbers. For example, E11 is the number of sentences expected to occur simultaneously in a document. Assume that we want to calculate these numbers for the following case as follows:

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{export} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

$$\begin{aligned} I(U;C) &= \frac{49}{801{,}948} \log_2 \frac{801{,}948 \cdot 49}{(49+27{,}652)(49+141)} \\ &+ \frac{141}{801{,}948} \log_2 \frac{801{,}948 \cdot 141}{(141+774{,}106)(49+141)} \\ &+ \frac{27{,}652}{801{,}948} \log_2 \frac{801{,}948 \cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} \\ &+ \frac{774{,}106}{801{,}948} \log_2 \frac{801{,}948 \cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)} \\ &\approx 0.0001105 \end{aligned}$$

$$E_{11} = N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N} =$$

$$N \times \frac{49 + 141}{N} \times \frac{49 + 27652}{N} \cong 606$$

N is the number of whole documents. Now we can obtain other N_{tc}, E_{tc} values.

	$e_{poultry} = 1$	$e_{poultry} = 0$
$e_{export} = 1$	$N_{11} = 49 \quad E_{11} \approx 6.6$	$N_{10} = 27{,}652 \quad E_{10} \approx 27{,}694.4$
$e_{export} = 0$	$N_{01} = 141 \quad E_{01} \approx 183.4$	$N_{00} = 774{,}106 \quad E_{00} \approx 774{,}063.6$

Replacing table numbers in formula X^2, we will find number 284.

$$X^2(D,t,c) = \sum_{e_t \in \{\varnothing,1\}} \sum_{e_c \in \{\varnothing,1\}} \frac{({}^{N}e_t e_c - {}^{E}e_t e_c)^2}{{}^{E}e_t e_c} \cong 284$$

X^2 shows the difference between expected observation numbers (E) and observed numbers (N). N value for X^2 represents the enhancement of independence theory. In this sense, there is a great interval between expected value and observed value. In the above example, the value $X^2(284)$ is larger than threshold 10/83. Therefore, one can say that for t, c cases, the chance level is in the row belonging to 0.001 (according to table1). If there is a great similarity between t and c, we can expect lower value for X^2. In other words, X^2 value represents dependency or similarity between t and c so that the smaller the value, the greater the similarity. Another method for calculating X^2 is as follows:

$$X^2(D,t,c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10} + N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) + (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$

TABLE I Acceptance value

P	X^2 (Critical value)
0.1	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

XI. CREATING CORPUS WITH AN INTELLIGENT AGENT

A parallel text corpus is referred to as a collection of texts which exist as parallel translations of sentences in two or more languages, and are considered as the most important prerequisite for many multilingual studies such as development of multilingual vocabularies and machine translation systems based on text corpora.

However, it is difficult to find a parallel text corpus for most languages except for a few pairs. Moreover, comparable text corpora are regarded by many researchers as a source for translation studies.

Comparable text corpora consist of pairs of documents in which texts in the same topic appear in different languages but they aren't necessarily translations of one another. They aren't parallel (sentence by sentence translation) so they don't consist of pairs of bilingual sentences with precise translation but include a level of parallelism in some degrees.

For example, there may be parallelism at the level of words, phrases, terms and sentences depending on corpus characteristics.

High quality parallel corpora with are seldom compiled for the pair of Persian and English languages or can't be accessed for research and commercial use due to copyright limitations. Thus, one of the priorities in developing a robust English-Persian machine translation system based on text corpora is to construct a parallel text corpus for these two languages.

Since the methodology of this investigation is based on text corpora and the main objective of this study is to extract correspondent chunks in Persian and English languages, it is necessary to make and use monolingual English and Persian text corpora separately. In another study conducted by the researcher, a Persian text corpus consisting of more than 2,640,000 sentences and 149,000,000 words was produced. Moreover, a bilingual parallel corpus consisting of more than 120000 sentences in English and Persian has been compiled. In this study, these two monolingual and bilingual text corpora were completed and expanded using software intelligent agents.

Thus, before explaining the way of using these intelligent agents in this study, the agents and software structure implemented in gathering and completing monolingual text corpus will have to be introduced.

Intelligent Agents (I-agents or IAs) are computer programs that may assist the user in computer applications. I-agents may be on the Internet, or can be on mobile wireless architectures. In the context of this research, however, the tasks that we are primarily concerned with include reading, filtering sorting, and maintaining information [6].

An agent may be a person, a machine, a software code segment or any other things which can be defined as any thing with authority. Agent is anything that can understand environment by its sensors and affect environment by its performance organs. For example, a human agent has sensory organs like eye, ear, etc. and performance organs like leg, hand, tongue, etc. A robotic agent uses different mechanical engines to affect environment and uses camera and other equipments to understand environment [6,8].

The main characteristic of an agent is its autonomy. An agent may act autonomously and control its internal mode. An agent is a software system and can work on dynamic and static environments in an autonomous and flexible manner. Flexibility means that agent may be passive, self-centered and social.

Passive agents don't have pattern, thought or memory, and only react against environment changes or messages received from other agents.

Self-centered agents have other additional control layers like design, logic, learning and even memory to react against environment and other agents. These agents provide new objectives with arguing and studying new conditions and act in this direction. In other word, these kinds of agents act in favor of themselves. Social agents act parallel to other agents and in a social environment.

An agent is a software system which has characteristics like "understanding environment and doing changes in it", autonomy, consistency (respond to environment changes) and being social.

An agent always gains prior knowledge from the environment, prior experiences by which it can learn, an objective which should access to, and information and observations about itself and surrounding environment and doing an action.

Artificial intelligence researchers believe that an agent is a computerized system which has some human characteristics such as knowledge, belief, intention and obligation in addition to the above general characteristics. From this viewpoint, agents have the following additional characteristics [8]:

- Portability: agent can move in an electronic network.
- Truth: agent doesn't transfer improper information intentionally.
- Benevolence: the objectives of agents do not vary, and any agent attempts to do only its assigned task.
- Rationality: an agent behaves in such a way that its objectives may be achieved to.

Every agent consists of two sections, program and architecture. Program is a function which implements agent behaviour, in other words, it is responsible for mapping from agent perceptions to a special behaviour. One of the artificial intelligence tasks is to design agents programs. Computational hardware on which agent program is conducted is called "agent architecture". Architecture can be a simple computer and include special hardware equipments like voice and video processing instruments. Hardware should also include software inserted between computer and agent program and provide the possibility of high level programming [6].

Architecture exposes observations to agent by sensors, executes programs and transfers agent behavior to environment and other agents by performance organs.

There are three major approaches to building agents for the WWW. The first approach is to integrate I-agents into existing search engine programs. The agent follows predefined rules which are employed in its filtering decisions. Making use of this approach has many advantages.

The second approach is a rule-based one. With this approach, the information about the application is given to an agent. Knowledge engineers are required to collect the required rules and knowledge for the agent. The third approach is a training one. In this approach the agent is trained to learn the preferences and actions of its user.

In this research, I-agents are used to retrieve data and information from the WWW. The Java programming language was used to create an I-agent. The I-agent actively develops searches out of desired data and information on the Web, and filters out unwanted data and information before delivering the results [11,12].

The simulation model assumes that, first, a connection to the WWW via a protocol, such as HTTP (Hypertext Transport Protocol), is done. Next, a URL (Universal Resource Locator) object class can be easily created. The URL class displays a pointer referring to a "resource" on the WWW. A resource can be something as simple as a file or a directory, or it can be a reference to a more complicated object, such as a query result via a database or a search engine.

Filtering and retrieving of information from the WWW using the I-agent, with the use of evolutionary computing and fuzzy logic according to keywords provided by the user is described as follows:

Phase 1:

Select the required search engine(s), such as Google, HotBot , Infoseek , Northernlight, etc.

Combine the keywords (k1, k2, …, kn) given by the user in a form understandable to the search engine(s) and submit the keywords to the i-agent.

Obtain the results of the search from the selected search engine(s). The host machine (of the search engine) returns the requested information and data with no specific format or acknowledgment.

Phase 2:

The i-agent program then calls its routines to identify all related URLs obtained from search engine(s) and inserts them into a temporary list (only the first 600 URLs returned are chosen) referred to as "TempList".

For each URL in the TempList, the following tasks are performed.

Once all URLs are retrieved, the initialized class generates zero (of the evolutionary computing population) using the supplied URL by the i-agent (Given an URL address from TempList , connecting to that web page).

A. **Agent's start point**

In order to determine an initial point to move software agent to, hamshahri newspaper website (Hamshahrionline.com) was considered as input source. Since genres of texts should be specified in a structure designed for Persian corpus, agent moves each time from a cluster (politics, religion, art, etc). After reading each text existing in the web pages related to determined section, the text enters into the corpus and the type of sentence becoms clear. Since each web page contains other relevant web pages, the relevant agent enters into the next section link by reading HTML and XML structure of that web page and the text contents of that section are also added to the corpus and this will continue until reading the whole site.

After accomplishing the work in this part and observing experimentally the retrieved texts, errors in software agent are specified and resolved. It is necessary to note that this agent acts like a Robot and after releasing in web environment it tries to gather the required data and send the retrieved documents to the text corpus.

Then, a list of valid sites containing relatively formal texts which are given to software agent as "Driving list" is selected in order to increase the textual corpus content. Then, the command of movement towards the agent issued using GO command and the software reads, records and moves to the next web page and web site according to a predefined program.

This act was performed completely automatic and without help of hand in terms of programming conducted in software agent. Different kinds of classified texts in Persian corpus are categorized into the following groups as politics, medicine, literature, sport, art, law, religion, science, events, economy and others.

Since different web sites used various font standards to store their web pages, the whole text is changed into Unicode (utf-8) format after retrieval.

While reading web pages, we faced a large amount of unrelated data such as images, tables and links to other pages that are omitted after specifying their type. For this purpose, first we transform read structure in XML standard form with specified tags and than insert it in the text corpus.

Due to the large amount of information gathered, the SQL Server 2005 was selected as software agent database in order to increase the speed of retrieving the corpus information. The main reason for selecting this database server is its high capability in processing SQL Queries. Moreover, compatibility of such database with most programming languages such as C# as well as database storage structure on clustered computer hardware is another reason for this choice.

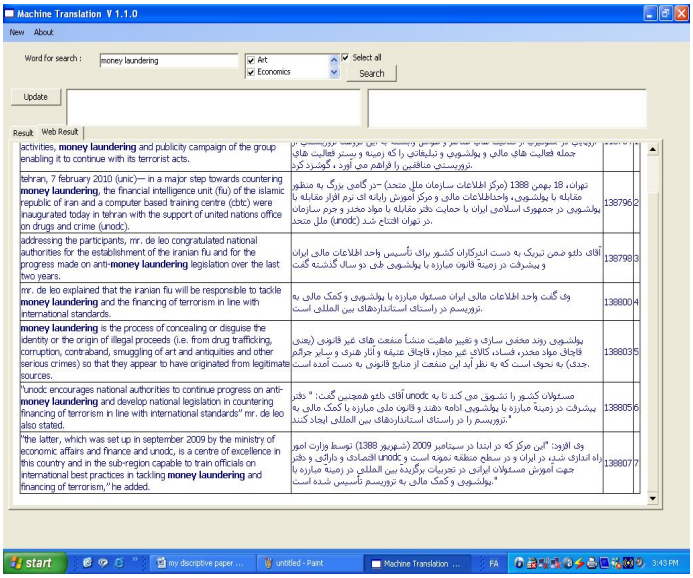

Fig. 1 Retrieved record for "money laundering" phrase.

B. **Software algorithm**

In order to obtain all chunks in source and target languages, software functions as follows:

At first, the whole structure of a sentence is extracted and then a forward movement value G is specified. The forward movement value is a number which can specify the maximum words which can be existed in a valid chunk.

In this project, G is considered 4 based on linguistic experience. Then, all possible compounds are identified from the hidden chunks in the sentence using computerized software.

In order to determine valid chunks and omit additional data, we classify each chunk. As explained in the section devoted to classification of texts in order to calculate X^2, if we measure dependency or relevancy rate in each expression, we will be able to decide on elimination or validity of chunks obtained.

For this purpose, we can use X^2 method explained in prior sections. This method was selected since we can determine dependency rate of two words c and d in a corpus using (d , c) X^2 formula. One can carefully observe that in this method, two words c and d occurred in all sentences in the corpus and frequency of different compounds containing occurrence or non-occurrence of each one or combination of them are measured. Then, dependency of each one is obtained using X^2 method.

$$X^2(D,t,c)=\frac{(N_{11}+N_{10}+N_{01}+N_{00})\times(N_{11}N_{00}-N_{10}+N_{01})^2}{(N_{11}+N_{01})\times(N_{11}+N_{10})+(N_{10}+N_{00})\times(N_{01}+N_{00})}$$

Consider the two following sentences from the English-Persian bilingual corpus:

چه تعداد سرباز را می تواند به خدمت سربازی درآورد؟-

-How many soldiers can it put into uniform?

After calculating X^2 value for each multiple compound, the following table was obtained:

TABLE III

X^2	Chunk	X^2	Chunk	X^2	Chunk
8.241	را می تواند به	72.023	تعداد سرباز را می تواند	6.02	**چه تعداد**
.	...	3.453	سرباز را	10.432	**چه تعداد سرباز**
.	....	8.0032	سرباز را می تواند	254.24	**چه تعداد سرباز را**
2.105	خدمت سربازی	11.25	سرباز را می تواند به	12.023	**تعداد سرباز**
1.231	خدمت سربازی درآورد	13.431	را می تواند	33.1	**تعداد سرباز را**

Now, in order to select desired compounds, a threshold value of 6.63 is selected based on the calculation of different values by researcher.

If X^2 or AS value calculated for each compound is smaller than threshold value of 6.63, we can accept dependency between words in that combination and reject values higher than threshold limit. This procedure is continued until all chunks in Persian corpus are identified and a new database from Persian chunks is formed.

Now, we try to find all records containing the relevant chunks using an English-Persian bilingual corpus in order to determine English equivalents of chunks obtained. For example, we consider multiple compounds « خدمت سربازی درآورد » “put into uniform” and extract the relevant records based on the method previously discussed from bilingual corpus.

Now we perform following routine for records obtained and then try to calculate and examine the result.

C. **Equalization algorithm**

The consequent stages in the equalization algorithm are as follows:

1- extract all two or four-fold compounds from English corresponding sentence.
2- calculate AS or X^2 value for each compound using English unilingual corpus.
3- keep compounds with AS values lower than 6.63 and omit the remainder.
4- using Persian compounds obtained from prior section and English compounds obtained from prior stage, the same tabular algorithm like below table is made for all different compounds as follows.
5- AS value related to table compounds is calculated and values lower than threshold are kept and we notice that the most value of AS is equal to relevant compound and it is “put into uniform”.

TABLE IIIII

English Chunk	Persian Chunk	AS	English Chunk	Persian Chunk	AS
How many	خدمت سربازی درآورد	194.12	.....	خدمت سربازی درآورد	
How many soldiers	خدمت سربازی درآورد	121.1	.....	خدمت سربازی درآورد	
Many soldiers	خدمت سربازی درآورد	133.99	.....	خدمت سربازی درآورد	
Many soldiers can	خدمت سربازی درآورد	..	.....	خدمت سربازی درآورد	
Soldiers can	خدمت سربازی درآورد	..	.....	خدمت سربازی درآورد	
Soldiers can it	خدمت سربازی درآورد	..	Put into	خدمت سربازی درآورد	**20.131**
Can it	**خدمت سربازی درآورد**	**..**	**Put into uniform**	**خدمت سربازی درآورد**	**4.2**

XII. ANSWER TO RESEARCH QUESTIONS

Considering above mentioned topics and the implemented laboratorial specimen, we should state in answering to the first question that we can use dictionary in compound words translation just as a validating measure. In this study , most of words have been translated using clustering texts and suggested algorithms. In respect of second question it should be stated that we can complete a corpus from webs textual data like an internet robot .

This can be done because of displacement ability of intelligent agent and independence of execution field .In answering to the third question, in which the use of multi-language corpus is considered in cross-language retrieval, we can say that it is necessary to use an independent database along with dictionary because Persian words are complex. In fact, this database is the language corpus which contains sentences and their equivalent translation in target language. We can use this corpus as a validating measure after translating word with dictionary.

XIII. CONCLUSIONS

This study was an attempt to present a statistical classification method using two text corpora (monolingual and bilingual) so that the dependency rate of phrases was determined and the ambiguity in translation of multiple phrases (chunk) was removed.

One of the main issues in retrieving linguistic information from the web or any other sort of database is that most users find it difficult to construct and retrieve complex queries in the language other than their own one. In most cases they set to translate the queries which are mainly in the form of phrases, collocations or simply chunks. The problem will be duplicated when the users' queries are in a specific domain while the existing translation of such queries falls into the general domain. The method presented in this study is the first attempt towards solving all mentioned problems in bilingual English-Persian retrieval systems.

In this study, in addition to complete monolingual and bilingual text corpora, a database from all chunks was made, and their translation was determined based on the information existing in corpora.

In this method, a candidate translation is specified using X^2 formula which shows the dependency rate between a chunk and its possible translation in an alternative language based on sentences in corpora as well as making use of the threshold table or critical point.

One of the consequences of implementing such a project is to increase the accuracy and precision of cross-language information retrieval systems in search engines in which word combination is accessible through database and corpus. Since a bilingual corpus of chunks in English and Persian with a very rich database was produced at the end of project, it can be predicted that using knowledge obtained from this study, products such as a corpus-based dictionary as well as a translation memory system are to be developed in separate investigations.

As the model presented in this study does not depend on specific properties of the languages involved in information retrieval, that is, it's a totally language-independent model whether in terms of linguistic or computational models, the algorithms presented in this study can be well applied to any other pair of languages.

In future studies, researchers intend to use made corpus as knowledge recourse and a new dictionary. Thus, data searching methods are suggested in order to complete single and binary corpus. In order to the speed of computation, we can also increase the speed of clustering words and their equivalent making by changing the method of computing AS .

REFERENCES

Carlson, C. N. 2004 ."Information Overload, Retrieval Strategies and Internet user Empowerment".

Chen, H .H .2002 .Chinese information extraction techniques .Presented at the SSIMIP, Singapore.

Douglas O. W., and B. J .Dorr .1996 .A survey of multilingual text retrieval .*Technical Report UMIACS-TR-96-19, Institute for Advanced Computer Studies,* University of Maryland, College Park , MD, USA .Xxii , 522, 528.

Hull, D .and G .Grefenstette .1996 .Querying Across Languages; A Dictionary –Based Approach to Multilingual Information Retrieval .In Proceedings of the 19th Annual International ACM Sigir, 49-57 .Zurich, Switzerland.

Lewis, D. and M. Ringuette. 1994. A comparison of two learning algorithms for text categorization. In Proc . SDAIR , pp . 81–93. 286, 526, 529.

Luck, M., and L. Padgham, (eds.). 2008 .Agent Oriented Software Engineering VIII :*The 8th International Workshop on Agent Oriented Software Engineering, AOSE 2007*, Honolulu, HI, May 14, Revised Selected Papers LNCS 495 .(Berlin, Germany :Springer Verlag

Miller, G.; R. Beckwith; C. Fellbaum; D.Gross, and K. Miller 1993. *Introduction to WordNet :An On-line Lexical Database*

Mohammadian, M. 2004 *Intelligent Agents for Data Mining and Information Retrieval*. Hershey :Idea Group Publishing.

Mosavi Miangah , T. 2008 .Automatic term extraction for cross-language information retrieval using a bilingual parallel corpus .*Proceedings of the 6th International Conference on Informatics and Systems)INFOS2008(*, PP .81-84, Cairo, Egypt.

Mosavi Miangah , T. 2009. Constructing a large-scale English-Persian Parallel Corpu" .META, 54 (1), pp .181-188.

Sihem, A. Y., and M. Lalmas . 2006 .XML search :Languages, INEX and Scoring .SIGMOD record 35(4): 16–23 .DOI : doi.acm.org/10.1145/1228268.1228271 .217, 519, 526.

Jensen, J. (2002). Using an intelligent agent to enhance search engine performance . Retrieved January 2002.