A NEW STATISTICAL MODEL
To Designing a Decision Support System

Morteza Zahedi, Ali Pouyan and Esmat Hejazi
Computer and Information Technology Department, Shahrood University of Technology, Shahrood, Iran

Keywords: Decision Support System, Technical Support Group, Statistical Pattern Recognition, Hidden Markov Model, Human Judgment, Expert Systems.

Abstract: In this paper we propose a new statistical approach to simulate a technical support center as a help desk for a web site which makes use of scientific documents and university protocols for the students and lecturers. In contrary to the existing statistical approaches which are modelled by general statistical graphs named Bayesian network or decision graph, we propose a statistical approach which can be used consistently in different domains and problem spaces without any need for a new designing regarding the new domain. Furthermore, the proposed statistical model which is trained by a set of training data collected from the experts in a special field is applicable to high-dimensional, large-sized, non-geometric-based data for decision making support.

1 INTRODUCTION

Corporations and companies often provide a help desk support to their customers in the sense of installation and usage of the products and troubleshooting the problems. Also, some schools offer classes in which they perform similar tasks as a help desk to help the students. In addition to the typical help desks, there are also many technical support forums freely available on the Internet, wherein expert and experienced users volunteer to help novices particularly in the field of computer programming and coding. As inside the companies, institutions and schools, employees and teachers need some information and technical support guides, there are also in-house help desks providing the same kind of help for employees, lecturers, or other internal associates only. It is very important for these services to be accessible 24-hours a day.

The various kinds of help desks introduced here are implemented via a toll-free telephone number or various online media such as website and/or e-mail.

Not only a help desk is very useful to support the costumers and offered services of a company to the costumers, but a help desk software can also be an extremely beneficial tool when it records the receiving questions and problems from the costumer side. The recorded information can be used to find, analyze, and eliminate common problems in an organization's computing environment. As a help desk communicates daily with numerous customers or employees, this gives the help desk the ability to monitor the user environment for issues from technical problems to user preferences and satisfactions. Such information gathered at the help desk is very valuable in planning and preparation to other units in the departments such as sales and product development (Middleton, 1996).

A very important problem for many help desks is to be strictly rostered. Time is an important parameter in such help desks to perform some tasks such as following up problems, returning phone calls, and answering questions via e-mail. The incoming phone calls and receiving queries by email or recorded messages via web portals are random in nature, so a rostered help desk agent schedules should ensure that all analysts get time to follow up on calls, and also ensure that analysts are always available to take incoming phone calls.

Due to the time constraints to response the receiving questions and high variaty of problems that occur in the services, causing different queries from the user side, the help desk is often referred to as the “hell desk” by the desk staff who work there. Thus, a decision support system (DSS) which is able to give some advices to the desk staff or answering the queries on the desk automatically is very useful. It can be emplemented along many new upcoming

Zahedi M., Pouyan A. and Hejazi E. (2008).
A NEW STATISTICAL MODEL - To Designing a Decision Support System.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 524-530
DOI: 10.5220/0001704905240530

technical support organizations which offer comprehensive computer repair services or guiding the users of a web portal.

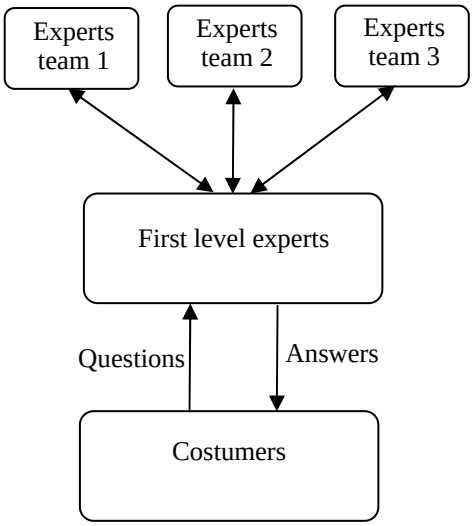

Figure 1: A two-level help-desk.

A typical help desk has two levels to handle different types of questions. It provides the users a central point to receive help on various problem issues. The user notifies the help desk of his or her issue, and the help desk issues a ticket including details of the problem. The first-level help desk is prepared to answer the general queries, the most commonly asked questions, or provide resolutions that often belong to the list of frequently asked questions (FAQ). The second level help desk consists of specialized teams ready to solve more complicated problems which are not solved in the first level. If the first level is able to solve the issue, the ticket is closed and updated with documentation of the solution to allow other help desk technicians to use it as a reference. If not, it will be dispatched to a second level where more experienced and expert people in a particular field are waiting for the queries. The queue manager will assign a ticket to one of the specialized teams based on the type of the issue. These comprehensive help desks need a team of experts in the first and second level to answer general questions and queries, and some specialists in the particular fields of the services, to work on-line 24 hours a day.

In this work we introduce a comprehensive help desk which uses the knowledge of the experts in a particular case. Not only can it help the desk staff as a decision support system by providing some advices, but it can also work alone instead of an expert team answering the questions and giveing advices to the visitors and users of a web portal automatically.

2 STATE-OF-THE-ART AND RESEARCH OBJECTIVES

In contrary to the expert systems implemented by knowledge base construction rules in the artificial intelligence (AI) discipline like genetic algorithms (Leu S. S., 2002, Turban, 2004), and knowledge representation methods (Michalewicz, Z.,2005, Bing Nan Li, 2008, and Ren-Jye Dzeng, 2007), the statistical methods aims at providing a rational decision in the context of probability theory and decision theory. The decision support systems which make rational decisions use collection of data gathered from the experts in a special field to construct statistical models.

It has been rather convincingly presented in numerous empirical researches that human judgment and decisions made by the experts are based on intuitive strategies. The intuitive strategies oppose to rational decision making methods which theoretically use reasoning rules. Empirical evidence and also several studies of expert performance in realistic settings show that experts and experienced people are more accurate than novices within their area of expertise, even though they are also liable to the same judgmental biases as novices and apparent errors and inconsistencies occur in their judgment. An informal review of the available evidence and literature review can be found in the book by (Robyn M. Dawes, 1988). Although the decisions made by heuristic methods are not based on optimal decision making rules and violates probability axioms by judgment biases, the intuitive strategies or judgmental heuristics help the experts and expert systems in the context of decision making by reducing the cognitive load. In an anthology edited by Kahneman, Slovic, and Tversky (Kahneman, 1982), there is a formal discussion of the most important research results along with experimental data.

In the general statistical approaches proposed for the decision support systems, very detailed analysis of domain tasks and information analysis theory is used (Bertin, 1983) to construct statistical models. Typically, the statistical approaches construct a Bayesian network or other kinds of decision graphs which is strictly dependent on a problem space (Dorner S., 2007). These methods focus on the special domain and problem space, analyze each node in the

problem space, identify each data item and information component (variable) for the nodes, the characteristics of the information components, and the relationships among the nodes. Finally, a conceptual database is used to store and retrieve the data collection gathered from the experts. The relationships among variables and rules that apply to problem-solving activities are described as a set of knowledge and form a conceptual knowledge base.

Also, some researchers have tried to represent huge data sets graphically. Computer hard disk usage is represented by development of TreeMaps, where huge data sets are involved and compressed (Shneiderman, 1992). Also, a punctuation graph is used to represent a technical document to allow the writer to detect potentially overly complex sentences, as well as to recognize familiar patterns (Perlman, 1983). The approaches using Bayesian networks or statistical graphs suffer from the limitation that exists in many business domains; the relationships among data are very complicated and cannot be presented with geometric structures such as hierarchies, linear, or networks. In other words, due to the nature of non-geometric data or non-spatial data, there is no obvious physical model that can be used to represent the data that humans can understand objectively.

Furthermore, in most business and management domains, problem-solving is overwhelming because of the large amount of complicated data, multiple complex relationships among data, and the negotiability of the constraints. Thus, in such systems including the data with complicated structure, it is difficult to construct a decision graph. Furthermore, general purpose representations are not easy to apply to a specific domain due to the complexity of data in different domains and sophisticated underlying functionality.

Another limitation of general statistical graphs, is that the data to be graphed must have controllable size or dimensions. In other words, the statistical approaches based on decision graphs cannot represent high-dimensional, large-sized, non-geometric-based data for decision-making support.

This research paper focuses on developing a research strategy for building a statistical model able to be used for non-geometric data that are massive in both size and dimensionality to help decision makers eventually to improve problem-solving performance or work alone instead of a group of experts. We will then apply the proposed statistical model to concrete a realistic domain to verify the effectiveness of the model. However, the proposed statistical model itself is domain-independent. It indicates the procedure of a decision support system as an automatic machine translation system which first maps receiving questions from the user-side into an answer from the desk-side. The final output of the proposed model is a set of most probable answers offering the desk staff supporting the entire human problem-solving process in a specific business domain. By employing a nearest neighbor (NN) classifier the best answer obtained from the statistical model can be chosen as an exact answer from the desk-side.

3 STATISTICAL MODELING OF THE DECISIONS

A simple view of decision making is that it is a problem of a choice among several alternatives. A somewhat more sophisticated view includes the process of constructing the alternatives, i.e. given a problem statement, developing a list of options. A complete picture includes a search for opportunities of decisions, i.e. discovering that there is a decision to be made. For instance, a manager of a company may face a choice in which the options are clear, e.g. the choice of a supplier from among all existing suppliers. There are a lot of anecdotal and some empirical evidence that structuring decision problems and identifying creative decision alternatives determine the ultimate quality of decisions. Decision support systems aim mainly at this broadest type of decision making, and in addition to supporting choice, they aid in modeling and analyzing systems, such as complex organizations, identifying decision opportunities, and structuring decision problems.

In other words a decision support system can be simplified to a machine translation (MT) system translating a source sentence to a target sentence.

Machine translation is a sub-field of computational linguistics, investigating the use of computer software to translate text or speech from one natural language to another. Simple machine translation methods perform simple substitution of words in one natural language for words in another. More complex translations may be attempted, by using corpus techniques in order to allow better handling of differences in linguistic typology, phrase recognition, and translation of idioms, as well as the isolation of anomalies. In order to improve accuracy of the MT methods, some research groups allow for customization by domain or profession (such as

weather reports) by limiting the scope of allowable substitutions.

Statistical machine translation is a kind of translation method trying to generate translations using statistical methods based on bilingual text corpora, such as the English-French record of the Canadian parliament and EUROPARL, the record of the European Parliament. By using such corpora, impressive results are obtained translating texts of a similar kind, but the scarceness of such corpora is still a critical problem for machine translation. Although the CANDIDE is the first statistical machine translation software from IBM, currently Google employs a statistical translation method improving their translation capabilities by inputting approximately 200 billion words from United Nations materials to train their system (Hutchins, W. John, 1992).

Although a typical MT system translates a source text into a target text in another language, it looks very similar to a decision support system which maps the user-side questions into the domain of desk-side answers. Machine translation is not always applied for a complete and accurate translation of texts. Sometimes a machine translation system is employed to perform a rough translation of a foreign language text, like a web page or news, which gives an idea of its contents. An MT machine is also applied for translation aid systems to help human translators. An inaccurate machine translation system can be used similarly as a decision support system.

In all these contexts, it is important to know when the system possibly made an error, and when one can be sure of obtaining a good translation. Since often a translation of a sentence as a whole is incorrect, but contains correct parts, the output of the MT system can provide the desk staff with a list of most probable answers. Also it is possible to employ a nearest neighbor classifier to find the most similar answer to the target sentence among the list of the answers collected from the experts in the particular field.

The statistical approach to machine translation has received growing interest over the last years since its introduction by the IBM research group in the early nineties. In various comparative evaluations, it has been proven to be competitive or superior to other traditional approaches. The translation quality achieved in restricted domains is relatively high. Examples include the domains of appointment scheduling, which was the scope of the project Verbmobil (W. Wahlster, 2000), or tourism which is used in the IWSLT evaluations (Y. Akiba., 2004). In recent years, more challenging tasks have been tackled in SMT research. The TC-STAR project (TCS, 2005), for example, deals with speech translation of the plenary sessions of the European Parliament. The domain and the vocabulary of these speeches are open.

The goal of machine translation is the automatic translation of a source language string $f_1^J = f_1 ... f_j ... f_J$ of words f_j into a target language string $e_1^I = e_1 \cdots e_i \cdots e_I$. In statistical machine translation (SMT), the translation is modeled as a decision process:

Given a source string f_1^J, the target string e_1^I with maximal posterior probability is determined:

$$\begin{aligned} \hat{e}_1^{\hat{I}} &= \arg\max_{I, e_1^I} \left\{ \Pr(e_1^I \mid f_1^J) \right\} \\ &= \arg\max_{I, e_1^I} \left\{ \frac{\Pr(f_1^J \mid e_1^I) . \Pr(e_1^I)}{\Pr(f_1^J)} \right\} \\ &= \arg\max_{I, e_1^I} \left\{ \Pr(f_1^J \mid e_1^I) . \Pr(e_1^I) \right\} \end{aligned} \quad (1)$$

Through this decomposition of the posterior probability $\Pr(e_1^I \mid f_1^J)$, two knowledge sources are obtained: the translation model $\Pr(f_1^J \mid e_1^I)$ and the language model $\Pr(e_1^I)$. Both of them can be modeled independently of each other. The translation model is responsible of linking the source string f_1^J and the target string e_1^I, i.e. for capturing the semantics of the sentence. The target language model $\Pr(e_1^I)$ assigns probabilities to target word sequences. It models the well-formedness or the syntax in the target language.

The probability of the source sentence, $\Pr(f_1^J)$, is usually omitted in the maximization because it does not affect the choice of the target word sequence. Nevertheless, it will be shown later that this probability is important for the methods suggested in this thesis.

The overall architecture of the statistical translation approach is depicted in figure 1.1.

The correspondence between the words in the source and the target string is described by alignments which can be viewed as mappings $a: j \to a_j \in \{1,...,i,...,I\}$ assigning a target position a_j to each source position j (Brown et al. 1993]. An artificial target position zero is introduced for mapping source words that do not have any equivalence in the target string. The alignment is introduced into the model as a hidden variable:

$$\Pr(f_1^J \mid e_1^I) = \sum_{a_1^J} \Pr(f_1^J, a_1^J \mid e_1^I) \quad (2)$$

Finally, using a nearest neighbour classifier, we find the most similar sentence among the target sentences to the translated sentence as the final choice of the decision support system.

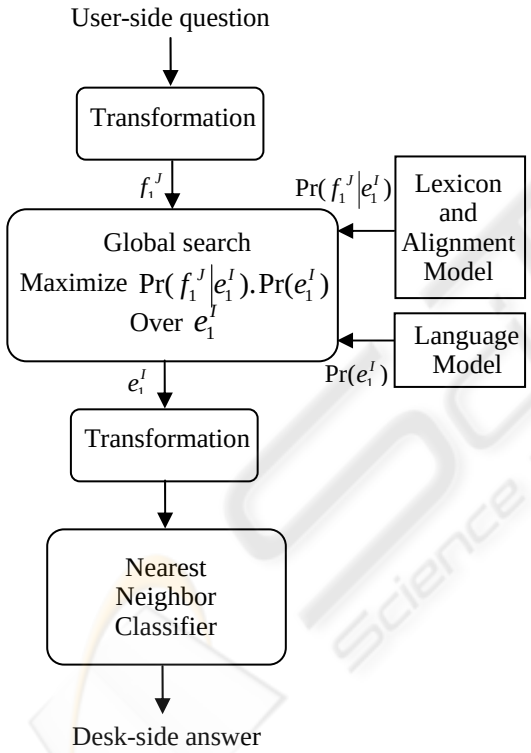

Figure 2: Architecture of the statistical approach for a technical support center.

4 PORSAJ TSG: A CASE STUDY

The PORSAJ is an on-line portal, sharing and selling electronic books, scientific documents, lecture notes, university examinations, and technical reports in Persian (http://www.porsaj.com). In this section we study the proposed statistical method to simulate the PORSAJ technical support group as a case study. In this section we introduce the technical support center of PORSAJ, and then we explain how the data set for the training, development and evaluation of the proposed statistical system is created.

4.1 The Technical Support Center

The Internet users consist of amateur visitors in the sense of smart, surfing through the pages to find needed information. When using an on-line portal, the visitors may face some problems which can be solved by asking a question from the technical staff of the web site. As it is explained before, a technical support center can be implemented by a web interface with the architecture of two-level help desk which at least a member from the technical support group is online and ready to answer the questions.

Currently, due to the high value of expenses the PORSAJ TSG can not work on-line for 24 hours a day. On the other hand, the PORSAJ visitors are transitive visitors who are linked from a search engine when searching for a document, they like to face an on-line comprehensive help desk. Thus a help desk equipped with a queuing system is not a good choice when the users have to wait for the answer of their questions some hours.

4.2 Data Set

The data set consists of 10,000 questions and sentences about the technical problems and some queries to get more information about the services of the website. These questions are collected from the help desk of the PORSAJ web site and also by filling the questionnaire forms by the web site visitors. The questions and sentences in the data set can be divided into three categories:

- Communicative messages like “Hello”, “Hi”, “Good morning”, etc.
- Questions about technical services and regarding problems.
- Some statements irrelevant to the subject of the services

These statements being received from the visitors should be answered by the on-line technical groups from the web site staff at the first level of the help desk. Any question or message received from the visitors is answered by a message from the technical group. The answers from the technical group are

reduced to 800 individual answers which are enough to answer the collected messages from the visitors.

The collected data set is divided into two parts; a training and an evaluation set. As the proposed statistical approach needs some parameters of the model to be trained we split the training set into two parts, a smaller training set and a development set. First, the statistical model is trained by the smaller training set and the parameters of the model are optimized based on the results obtained by using the model on the development set. Then, the optimized parameters are used to train the model by using the whole samples of the training and development set. Finally, the resulting model is tested on the evaluation set. The number of samples in the training, development, and evaluation set is listed in the Table 1.

Table 1: The statistics of the data set.

		User-side (questions)	Desk-side (answers)
Train	# Sentences	8000	800
	# Run. Words	90423	11096
	Vocab. Size	576	186
	# Singletons	142	98
Development	# Sentences	1000	442
	# Run. Words	8937	3820
	Vocab. Size	248	82
Evaluation	# Sentences	1000	800
	# Run. Words	9022	11096
	Vocab. Size	308	186

5 EXPERIMENTAL RESULTS

In order to evaluate the proposed method, we do the experiments on the data set which is introduced in the former section. First, we train the hidden Markov models for the lexicon and alignment models and also the language model by using the training part of the database. Then, the parameters of the HMMs like time distortion penalty, and language model scales are optimized on the development set. Finally, we test the resulting model of the DSS on the evaluation part of the data set. The results of the classifier on the development and evaluation set are listed in Table 2.

Table 2: The error rate of the DSS on the development and evaluation set.

	Sentence error rate
Development	8.2%
Evaluation	12.8%

The results show that the help desk can work alone with an accepted rate of correct answers. However, the results rely on training the data strongly and for some applications which the vocabulary size of the training set is very big, we expect to have more sentences in the training part in order to help the construction of the statistical model. Nevertheless, if we can not obtain an acceptable rate of right answers for any other applications, the model can be used as a DSS to help the technical support staff instead of working alone.

6 CONCLUSIONS

In this paper, a decision support system is suggested to be simplified as a machine translation system. In contrary to the general statistical graphs which suffer from some constraints like controllable size or dimensions of the data, The proposed approach focuses on developing a research strategy for building a domain-independent statistical model capable to be used for non-geometric data that are massive in both size and dimensionality to help decision makers eventually to improve problem-solving performance or work alone instead of a group of experts. We will then apply the proposed statistical model to concrete a realistic domain to verify the effectiveness of the model. The experimental results show that the statistical model can work with an acceptable rate of correct answers from the desk side.

ACKNOWLEDGEMENTS

We are very thankful for the Information and Communication Treasure Company (ICT Co.) and Shahrood University of Technology (SUT) funding

this research. Also we appreciate the efforts of the students of the Computer and Information Technology Department at the SUT.

REFERENCES

Bing Nan Li, Ming Chui Dong and Sam Chao, 2008, *On decision making support in blood bank information systems*, Expert Systems with Applications: An International Journal, ELSEVIER, ScienceDirect, Vol. 34 , No. 2, pp. 1522-1532.

Ren-Jye Dzeng and Hsin-Yun Lee, 2007, *Activity and value orientated decision support for the development planning of a theme park*, Expert Systems with Applications: An International Journal, ELSEVIER, ScienceDirect, Vol. 33, pp. 923-935.

Dorner S., Shi J, Swayne D., 2007, *Multi-Objective Modelling and Decision Support Using a Bayesian Network Approximation to a Non-point Source Pollution Model*, Environmental Modelling & Software, Vol. 22, pp. 211-222.

TCS, 2005, *TC-STAR - Technology and Corpora for Speech to Speech Translation*, Integrated project TC-STAR (IST-2002-FP6-506738) funded by the European

Commission. http://www.tc-star.org/.

Michalewicz, Z.; Schmidt, M.; Michalewicz, M.; Chiriac, C., 2005, *Case study: an intelligent decision support system*, Intelligent Systems, IEEE Vol. 20, Issue 4, pp. 44-49.

Turban, 2004, *Decision Support System and Intelligent Systems*, 6th Edition, Millan, 2004.

Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, J. Tsujii, 2004, *Overview of the IWSLT04 Evaluation Campaign*. In Proc. of the Int. Workshop on Spoken Language Translation (IWSLT), pp. 1–12, Kyoto, Japan.

Leu S. S., and Hwang, S. T., 2002, *GA-based resources-constrained flowshop scheduling model for mixed precast production*. Automation in Cnstruction, Vol. 11, No. 4, pp. 439-452.

Leu S. S., and Hwang, S. T., 2002, *GA-based resources-constrained flowshop scheduling model for mixed precast production*. Automation in Cnstruction, Vol. 11, No. 4, pp. 439-452.

W. Wahlster, 2000, *editor: Verbmobil: Foundations of speech-to-speech translations*. Springer Verlag, Berlin, Germany.

Middleton, I., 1996. *Key Factors in Help Desk Success (An analysis of areas critical to help desk development and functionality.)* In British Library R&D Report 6247, The British Library.

Hutchins, W. John, and Harold L. Somers, 1992. *An Introduction to Machine Translation,* London: Academic Press, Press. ISBN 0-12-362830-X.

Shneiderman, Ben, 1992. *Tree Visualization with TreeMaps: A 2-D Space-filling Approach,* ACM Transaction on Graphics, Vol. 11, No. 1, pp. 92-99.

Robyn M. Dawes, 1988. *Rational Choice in an Uncertain Choice*, Hartcourt Brace Jovanovich, Publishers.

Bertin, Jacques, 1983. *Semilogy of Graphics.* Translated by William J. Berg, At The University of Wisconsin.

Perlman, G. and T.D. Erickson, 1983. *Graphical Abstractions of Technical Documents*, In Visible Language, Vol. 23, No. 4.

Daniel Kahneman, Paul Slovic, and Amos Tversky, 1982. *Judgment under uncertainty: Heuristics and biases*, Cambridge University Press.