Toward a Universal Translator of Verbal Probabilities

Tzur M. Karelitz[1], Mandeep K. Dhami[2], David V. Budescu[1] and Thomas S. Wallsten[3]

[1]University of Illinois at Urbana-Champaign
603 E. Daniel St. Champaign,
IL 61820
karelitz@s.psych.uiuc.edu

[2] University of Victoria
P.O. Box 3050
Victoria, BC V8W 3P5

[3] University of Maryland
College Park,
MD 20742-4411

Abstract

When forecasters and decision-makers use different phrases to refer to the same event, there is opportunity for errors in communication.□In an effort to facilitate the communication process, we investigated various ways of "translating" a forecaster's verbal probabilities to a decision-maker's probability phrases.□We describe a blueprint for a general translator of verbal probabilities and report results from two empirical studies.□The results support the proposed methods and document the beneficial effects of two, relatively simple translation methods.

Introduction

People often use probability phrases to communicate different levels of uncertainty. Decision-makers rely on probability estimates given by human advisors or computer expert systems. Typically, these systems deal with uncertainty by using a fixed set of probability phrases. However, the meaning of phrases can vary from one person to the next. An analyst predicting that a plan has a good chance of succeeding might have in mind a 60%-70% probability of success. An executive deciding whether to implement the plan might interpret this good chance as an 80%-90% probability, thus leading her to make a different decision than she would have had she correctly understood the intended meaning. We refer to such misunderstandings as verbal-uncertainty errors. Although the consequences of such errors can be quite dramatic, the use of verbal probabilities is very common in the business, financial, judicial, medical, and political domains, and in computerized expert systems.

Erev and Cohen (1990), and Wallsten et al. (1993) demonstrated that people often prefer expressing uncertainty verbally rather than numerically. Many studies have shown that people have different lexicons for expressing uncertainty (Budescu, Weinberg, and Wallsten 1988; Erev and Cohen 1990; Zwick and Wallsten 1989) and that individuals comprehend verbal probabilities in different ways (e.g., Beyth-Marom 1982; Budescu and Wallsten 1985; Clarke et al. 1992; Mullet and Rivet 1991). Some of the latter studies also report that variability between individuals is greater than variability within individuals (across multiple replications) when assessing the same phrases. Artificial intelligence researchers have incorporated probability phrases identified in the psychological literature. For example, in an expert system named RUM (Reasoning with Uncertainty Module), Bonissone, Gans and Decker (1987), suggested using nine phrases from Beyth-Marom's (1982) list of verbal probabilities. López de Mántaras, Meseguer, Sanz, Sierra and Verdabuer (1988) described a medical diagnostic expert system (called MILORD) that encompasses nine probability phrases. We argue that computer-aided decision-making may also result in verbal-uncertainty errors, and we explore ways to minimize the effect of these problems.

The problem we identify in uncertainties communication can be symbolized with a communication channel model:

Sender → Encoder → Channel → Decoder → Receiver

The sender and receiver are the advisor and decision-maker, respectively. The encoder and decoder are the mental systems that produce and interpret words. The channel is some communication tool such as a written report, telephone, or computer communication. Usually the channel is considered the major source of noise in the system. However, in this model the errors originate in the encoding-decoding devices. In general, for communication systems to be effective, the decoder must be able to decode the encoded messages but, as mentioned earlier, people interpret probability phrases in different ways, thus, often times, the decoded meaning of a phrase is not the same as the pre-encoded meaning.

The primary goal of our research is to develop a method for translating one person's lexicon to another's, in order to facilitate communication by means of verbal probabilities. We suggest using the channel component as a translation device that calibrates the encoded message to the decoder capabilities. Since computers have become an integral part of daily communication between advisors and decision-makers, implementation of such a system is practical and feasible.

The issue of translation is not trivial as it involves establishing empirical criteria for assessing similarity of meaning across individuals and determining methods for selecting phrases that are the most similar to each other. We approach the problem by establishing two different, but related sets of similarity criteria (we illustrate one set of criteria in Experiment 1 and the other in Experiment 2), and by testing which of several translation methods best selects most-similar phrases. Indeed, both studies demonstrate the effectiveness of the same methods. The purpose of this paper is to summarize these methods as well as the empirical work supporting them to date.

Membership functions and phrase ranking

We used two classes of methods for matching phrases across lexicons. One was simply to have respondents rank-order the phrases in their selected lexicon from least probable to most probable, and we selected phrases that matched most closely in rank. The other class relied on membership functions (Zadeh 1965) to represent phrase meaning (Rapoport, Wallsten, and Cox 1987; Wallsten et al. 1986), and we selected phrases that most closely matched on some membership-function index. In this sense, each probability phrase is treated as a vague concept whose members are the numerical probabilities in the [0,1] range. Membership functions (MF), which may vary from one person to another, assign a number to each value on the probability scale that represents its degree of membership in the concept defined by the phrase. Degree of membership varies from 0 (probabilities definitely not in the concept) to 1 (perfect prototypes of the concept), but is not to be interpreted as a second-order probability.

Blueprints for our verbal probability translator and validating experiments

The components of our verbal probabilities translator are:

- Each participant chooses a set of verbal expressions that covers the full range of the probability scale [0,1] and that includes their *personal* lexicon. Note that different individuals might have different phrases and even different number of phrases in their lexicons.
- Each participant provides characteristic feature for each phrase. Candidate features are the rank order of the phrase in the set, and the phrase's MF (from which various indices, such as the probability at which the MF peaks, can be extracted). This information is used to create translation tables that match the meanings of phrases from the lexicons of each distinct pair of participants.

Naturally, these components are part of our validation studies but, in addition, we include the following items:

- Each participant provides verbal and (possibly at a different point in time) numerical judgments of the likelihoods of a large number of events (common to all members of the community) in a given domain.
- Two phrases are considered similar to the extent that they are used to describe the same events or the events to which they are applied receive the same numerical estimates. This determination relies on the assumption that the verbal and numerical estimates of an event are intended to convey the same degree of uncertainty.
- Indices of phrase similarity estimated from the MFs or rank orders are compared to those calculated from the numerical judgments, and to those obtained from unaided verbal communication to determine their relative performance, and the one that does best overall.

Overview of the methods and analyses

We highlight key results of two experiments designed to test the translation systems. The analysis of any such experiment involves:

- Choosing a set of translation methods based on the features of the verbal probabilities, such as their likelihood ranking, their MFs peaks, and the shapes of their MFs.
- Creating translation tables for each method and for every possible pair of participants, by matching each person's phrases to the phrases used by their partners.
- Choosing a criterion (or set of criteria) of effective communication. It is possible to define effective communication in many ways including the similarity between the sets of events to which participants apply the same phrases, the level of agreement between participants in assignment of phrases to events, or the improvement in agreement relative to some baseline (e.g., chance). These criteria enable us to evaluate the absolute and relative efficiency of the various translations methods.
- Measuring the effectiveness of numerical and verbal communication (without translation), and of each translation method. The former measures provide the lower and upper bound of effectiveness. Numerical communication should have fewer errors, since the communicator and the recipient use the same set of well-defined probability values. For the reasons outlined earlier, unaided verbal communication should be the least effective.

Translation methods

For simplicity, we will refer only to two persons: *i* and *j*. Sometimes *i* produces the verbal probabilities and *j* is the recipient, and sometimes the roles are reversed. The participants' lexicons may vary in size and content. Specific phrases from a certain person's lexicon are labeled w_{jm} (phrase *m* from person *j*) or w_{in} (phrase *n* from person *i*). Each translation method aims to match a phrase from person *j*'s lexicon to every phrase in person *i*'s lexicon, based on a specific criterion. Below is a brief description of the translation methods used in the experiments to follow.

ABSDEV (**ABS**olute **DEV**iation) - Calculate the mean absolute deviation (across all probabilities) between the MF of w_{in} and the MF of each phrase in person *j*'s lexicon. Choose w_{jm} - the phrase for which the sum is smallest.

PRO (**P**eak **R**ank **O**rder)- Choose w_{jm} such that its rank order (as inferred from the peaks of judge *j*'s MF's) matches the rank order of w_{in} (as inferred from the peaks of judge *i*'s MF's).

DPEAK (**D**ifference in **PEAK**s)- Choose w_{jm} such that the distance between the location of its peak and that of w_{im} is less than that of any other phrase in person *j*'s lexicon.

RANK Choose w_{jm} such that its rank (given by person *j*) is the closest (in absolute distance sense) to the rank of w_{in} (given by person *i*).

INTER For each w_{im}, determine the areas of intersection of its MF with the MFs of each phrase in person *j*'s lexicon. Determine the maximal MF value for each intersection and select as the matching w_{jn} that phrase for which the maximum is greatest.

An important feature of translation methods is symmetry of translation. Let i and j denote two participants. A method is symmetric if w_{im}, the m'th phrase used by person i is translated to w_{jn}, the n'th phrase used by person j and if w_{jn} is translated to w_{im}. A method is asymmetric if w_{im} is translated to w_{jn}, but w_{jn} is translated to w_{ik} and $w_{ik} \neq w_{im}$. when both individuals use the same number of phrases, only rank-based methods such as RANK or PRO are symmetric.

Experiment 1

Experiment 1 aimed to validate the use of translation methods for reducing errors in communication of uncertainties (here we present partial findings from a study by Karelitz and Budescu 2001). Our hypothesis is that communication errors caused by using verbal probabilities can be reduced by translating the verbal probability lexicon of person i to that of person j. We claim that the upper bound for the improvement of communication between two people by the use of translation is the level of error achieved if they were communicating exclusively using numerical probabilities. Conversely, the lower bound (baseline) for improvement is the un-translated verbal probabilities.

In this experiment, we presented a set of graphical stimuli to participants and collected their verbal and numerical probability judgments of the occurrence of the target event. We evaluated the quality of four translation methods by comparing the number and magnitude of errors in translated communication to the lower bound (un-translated Verbal Judgments - VJ), and to the upper bound (Numerical Judgments - NJ).

We predicted that the agreement indices would have the greatest values for numerical judgments, lesser agreement for the translated verbal judgments and the lowest agreement for the un-translated verbal judgments.

Method

Eighteen native English speakers volunteered to participate. The experiment consisted of three computerized tasks - (1) Selection and ranking of a verbal probability lexicon. (2) Elicitation of membership functions of the selected phrases. (3) Numerical and verbal likelihood estimation of graphically displayed events.

In the first task, participants were asked to create a list of 6-11 phrases by selecting combinations of phrases and semantic operators (e.g., modifiers, quantifiers, negations, intensifiers, etc.). They were instructed to select phrases they use in their daily lives that span the whole probability range. Three phrases in this list were pre-selected for all participants: *Certain*, *Even odds* and *Impossible*. After selecting the phrases, participants were asked to rank them in ascending order.

In the second task, the MFs of the selected phrases were elicited using the multi-stimuli method (Budescu, Karelitz, and Wallsten 2000). Each phrase appeared with a set of 11 probabilities ranging from 0 to 1, in intervals of .1. For each phrase, participants judged the degree to which the target phrase captured the intended meaning of each of the 11 probabilities. All judgments were made on a bounded scale ranging from '*not at all*' to '*absolutely*'.

In the third task, participants judged the likelihood of certain events occurring using numerical and verbal probabilities. On each trial, participants saw a circular target with shaded parts. Their task was to assess the likelihood that a dart aimed at the center of the circle will hit the shaded area. The numerical judgments were made by selecting a value from a list of 21 probabilities, ranging from 0 to 1 in intervals of .05. The verbal judgments were made by selecting and ranking up to four phrases from the participant lexicon. The stimuli were two sets of 19 circles presented in a random order to each participant. Each set covered all probabilities from 0 to 1 in increments of .05.

Results

We compared the verbal and numerical judgments of every stimulus for each participants dyad. Since each participant could have selected up to four phrases when making judgments, we used all possible combinations of phrases per pair to create two indices of co-assignment:

PIA- **P**roportion of **I**dentical **A**ssignments- the proportion of stimuli to which both participants assigned *the same* phrase, and

PMA- **P**roportion of **M**inimal **A**greement - the proportion of stimuli for which both participants assigned *at least one* common phrase.

The analysis was repeated using four methods of translation (i.e., methods 1 to 4 above). This was done by designating one participant in each pair as the communicator and the other as the recipient (and visa-versa). The phrases used by the communicator to describe each stimulus were translated into the recipient's lexicon, using each of the four methods. The comparison was done between the recipient's original phrases and the communicator's translated phrases. The average co-assignment indices for the four translation methods, the numeric judgments (NJ) and the un-translated verbal judgments (VJ) are presented in Table 1. Each of these values is based on 18 x 17 = 306 pairs of participants. Also included in Table 1 are the geometric means of the ratio between the agreement indices for each translation method and the agreement indices of the VJ. Ratios greater (smaller) than 1 indicate an improvement (deterioration) in communication as compared to the unaided verbal case.

Measure	ABSDEV	RANK	DPEAK	PRO	VJ	NJ	Ratio
PIA	0.21	**0.23**	0.21	0.18	0.05	0.29	4.2
PMA	0.67	0.78	0.76	**0.90**	0.26	0.68	3.0
Ratio	3.6	4.0	3.6	3.1			

Table 1 - Co-assignment indices and geometric mean of ratio between translation methods and VJ

It is clear that the four translation methods outperformed the baseline condition (VJ), and some outperformed the level of agreement achieved by NJ. In other words, all four translation methods reduced the level of error considerably according to our criteria.

Summary of Experiment 1

The main findings can be summarized in two points. First, translating one from person's lexicon to another's can reduce errors in communication of verbal uncertainties. Effective translation methods can be devised from MFs of probability phrases (PRO) or from their rank orderings (RANK). In fact, all translation methods outperformed the VJ, and most did better than the NJ under the more lenient criterion (PMA). The second conclusion is that different agreement indices favor different translation methods. The best method on one index was not always the best on the others (e.g., PRO on PMA and PIA).

We should mention some of the limitations of the present study. First, the translation methods were only tested in cases where uncertainty was quantifiable. There are many occasions when uncertainty is not measurable, because it is due to an internal lack of knowledge for example. Howell (1971) distinguished between aleatory and epistemic uncertainty. Individuals' understanding and use of verbal probabilities under these two conditions may differ. Second, although we asked participants to select their own lexicon, we provided them with pre-selected phrases. Our purpose was to provide them with some clear anchors and to make sure that their lexicon spanned the whole range, but it is possible that some people don't usually use the pre-selected phrases. These limitations are addressed in the next study.

Experiment 2

The second experiment (here we present some findings from a larger study by Dhami and Wallsten 2001) differs from the first in the following ways. We tested the methods of translation under conditions of both aleatory and epistemic uncertainty (which is not easily measurable). We did not pre-select any phrases, and although we allowed respondents to select their entire lexicon, we did fix the size of the lexicon to 7 phrases per person. It is not uncommon for people to spontaneously select only a handful of phrases (e.g., Budescu and Wallsten 1995). The translation methods we used were DPEAK, RANK, and INTER. Finally, we used a different empirical index of phrase similarity from those summarized in Table 1, which provides a means of comparing across aleatory and epistemic uncertainty. (We present only epistemic results.)

Method

Twenty-nine native English speakers volunteered to participate. The experiment comprised four computerized stages – (1) Selection and ranking of participants' verbal probability lexicons. (2) Verbal forecasts of the chances of a set of future events occurring. (3) Elicitation of MFs of the selected phrases. (4) Numerical forecasts of the chances of a set of future events occurring (same events used in stage 2).

In the first stage, participants were asked to select seven phrases they normally use to describe probabilities spanning the [0,1] interval. There were no pre-selected phrases. Participants then rank ordered their phrases.

In the second stage, participants forecasted the chances of 100 real-world events occurring in the future and of 100 aleatory events (similar to those used in Experiment 1). The real-world events were sampled from the domains of current affairs and politics, entertainment, sport, the University of Maryland, and science (e.g., "What are the chances that billiards will be included in the 2004 Olympics held in Athens?"). In the third stage, the MFs of the selected phrases were elicited as described in Experiment 1. Finally, the real-world events from stage 2 were repeated, and participants responded using numerical probabilities, on a 0 to 1 scale, with 0.1 intervals.

Results

In the remainder, we present the results for the real-world events. To determine which method best translates the meanings of phrases from one person's lexicon to another's, we first established a "gold standard." For this purpose, we inferred the numerical meanings of each participant's phrases from their usage in stages 2 and 4. Specifically, for each phrase and each participant, we estimated a cumulative frequency distribution of probabilities by taking all the events that were assigned that phrase in stage 2 and cumulating the probabilities that were assigned to them in stage 4. The gold standard was the maximal deviation between the distributions of pairs of phrases from different lexicons (the Kolmogorov-Smirnov or K-S statistic). Any effective translation method should therefore, accurately reproduce the gold standard.

The gold standard measure of inter-personal similarity of phrases per participant-pair was summarized in a respondent *i* by respondent *j* matrix, and the predictions of interpersonal similarity derived from DPEAK, INTER and RANK were summarized in similar matrices. Rank order correlations were computed between each method and the gold standard. As Table 2 shows, we found a high correlation between the gold standard, which was based on meanings of phrases as inferred from how people used them, and each of the three methods, which were all based on self-reported meanings of phrases. However, the RANK method was superior to other methods in capturing and translating the meanings of phrases across individuals.

	DPEAK		INTER		RANK	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Gold Standard	0.55	0.14	0.59	0.16	0.64	0.10

Table 2 - Rank correlations between predictions made by translation methods and the gold standard.

Summary of Experiment 2

All three translation methods predicted the gold standard well, thus demonstrating their efficacy in translating verbal probabilities across lexicons, under conditions of underlying epistemic uncertainty. The fact that the simple RANK method proved to be more effective in translation than the two methods based on MFs, suggests that MFs may not be required in a universal translator of verbal probabilities. However, because we restricted the number of phrases in participants' lexicons to 7, it is unclear how the methods perform under circumstances when participants have much smaller or much larger lexicons, and when participant-pairs have different size lexicons.

General Discussion

Our main hypothesis, that translation methods yielded better agreement indices than unaided verbal judgments, was supported by the data. The most important conclusion of the study is that translation methods are beneficial and *can* reduce errors in communication when people use different lexicons of uncertainty. These methods can be implemented into expert systems with relative ease. A decision aid that utilizes the user's verbal probability lexicon for communication, may improve the quality of the decision making process. Further research is needed to quantify the improvement in decision making rather than the agreement in judgmental tasks.

The fact that different indices of agreement quality tend to favor different methods makes the task of choosing the "best method" difficult. It is however, reassuring to know that relatively simple measures based on direct rankings or rankings inferred from MFs seem to work well in most cases. Further research on the meaning of these indices is needed, and a coherent selection process of the best method and the best criterion is essential. Another direction for prospective research is to explore the use of translation methods in evaluating probability phrases across contexts and different languages.

Acknowledgments

This work was supported by NSF grants SES-9975360 and SES-9975350.

References

Beyth-Marom, R. 1982. How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting* 1:257-269.

Bonissone, P. P., Gans, S. S., and Decker, K. S. 1987. RUM: A Layered Architecture for Reasoning with Uncertainty. In *Proceedings of IJCAI* 373-379. Milan, Italy.

Budescu, D. V., and Wallsten, T. S. 1985. Consistency in Interpretation of Probabilistic Phrases. *Organizational Behavior & Human Decision Processes* 36:391-405.

Budescu, D. V., Weinberg, S., and Wallsten, T. S. 1988. Decisions Based on Numerically and Verbally Expressed Uncertainties. *Journal of Experimental Psychology: Human Perception & Performance* 14:281-294.

Budescu, D. V., Karelitz, T. M., and Wallsten, T. S. 2000. Predicting the Directionality of Probability Words from Their Membership Functions. *Paper presented at the 41st Annual Meeting of the Psychonomic Society, New Orleans, LA.*

Clarke, V. A., Ruffin, C. L., Hill, D. J., and Beamen, A. L. 1992. Ratings of Orally Presented Verbal Expressions of Probability by a Heterogeneous Sample. *Journal of Applied Social Psychology* 22:638-656.

Erev, I., and Cohen, B. L. 1990. Verbal versus Numerical Probabilities: Efficiency, Biases, and the Preference Paradox. *Organizational Behavior & Human Decision Processes* 45:1-18.

Dhami, M. K., and Wallsten, T. S. 2001. Interpersonal Similarity in Uses of Linguistic Probabilities. *Paper presented at the 18th Conference on Subjective Probability, Utility and Decision Making, Amsterdam, The Netherlands.*

Howell, W. C. 1971. Uncertainty from Internal and External Sources: A Clear Case of Overconfidence. *Journal of Risk and Uncertainty* 4:5-28.

Karelitz, T.M. and Budescu, D.V. 2001. Evaluating Methods of Translating one Person's Verbal Probabilities to Another. *Paper presented at the annual meeting of the Society of Judgment and Decision Making. Orlando, FL.*

López de Mántaras, R., Meseguer, P., Sanz, F., Sierra, C., and Verdaguer, A. 1988. A Fuzzy Logic Approach to the Management of Linguistically Expressed Uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics* 18:144-151.

Mullet, E., and Rivet, I. 1991. Comprehension of Verbal Probability Expressions in Children and Adolescents. *Language & Communication* 11:217-225.

Rapoport, A., Wallsten, T. S., and Cox , J. A. 1987. Direct and Indirect Scaling of Membership Functions of Probability Phrases. *Mathematical Modeling* 9:397-417.

Wallsten, T. S., Budescu, D. V., Rapoport, A. Zwick, R., and Forsyth, B. 1986. Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General* 115:384-365.

Wallsten, T. S., Budescu, D. V., Zwick, R., and Kemp, S. M. 1993. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society.* 31:135-138.

Zadeh, L. A. 1965. Fuzzy Sets. *Information and Control* 8:338-353.

Zwick, R., and Wallsten, T. S. 1989. Combining Stochastic Uncertainty and Linguistic Inexactness: Theory and Experimental Evaluation of Four Fuzzy Probability Models. *International Journal of Man-Machine Studies* 30:69-111.