Published in final edited form as:
Mach Transl. 2014 March 1; 28(1): 1–17. doi:10.1007/s10590-013-9140-x.

A Conjoint Analysis Framework for Evaluating User Preferences in Machine Translation

Katrin Kirchhoff,
Department of Electrical Engineering, University of Washington, Seattle, WA, 98195, USA, Tel.: +1-206-616-5494, Fax: +1-206-543-3842, katrin@ee.washington.edu

Daniel Capurro, and
Department of Internal Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile, dcapurro@med.puc.cl

Anne M. Turner
Department of Health Services/Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, 98195, USA, amturner@u.washington.edu

Abstract

Despite much research on machine translation (MT) evaluation, there is surprisingly little work that directly measures users' intuitive or emotional preferences regarding different types of MT errors. However, the elicitation and modeling of user preferences is an important prerequisite for research on user adaptation and customization of MT engines. In this paper we explore the use of conjoint analysis as a formal quantitative framework to assess users' relative preferences for different types of translation errors. We apply our approach to the analysis of MT output from translating public health documents from English into Spanish. Our results indicate that word order errors are clearly the most dispreferred error type, followed by word sense, morphological, and function word errors. The conjoint analysis-based model is able to predict user preferences more accurately than a baseline model that chooses the translation with the fewest errors overall. Additionally we analyze the effect of using a crowd-sourced respondent population versus a sample of domain experts and observe that main preference effects are remarkably stable across the two samples.

Keywords

machine translation; evaluation; user modeling; preference elicitation

1 Introduction

Much recent work in the machine translation (MT) community has focused on evaluation metrics for MT output, especially on automatic evaluation metrics. However, there has been surprisingly little research that directly investigates which types of errors are intuitively disliked the most by actual users of MT. Although there is ample anecdotal evidence of users' reactions to MT errors, it is difficult to find formal, quantitative studies of how users perceive the severity of different translation errors, and what trade-offs they would make between different errors if they were given a choice. Having a formal framework for studying these questions is important for several reasons:

- Core MT technology has reached a certain level of maturity. MT engines are now routinely used by commercial and governmental/non-profit organizations, as well as by individual lay users. Rapid user adaptation and customization of MT engines are emerging as important future directions for MT research, and it is necessary to develop principled strategies for eliciting and modeling user preferences.
- In the case of limited research and development resources, knowledge of which errors users regard as serious vs. negligible may result in better prioritization of resources compared to addressing those errors that affect global performance as measured by automatic evaluation metrics.
- User preferences might sometimes diverge strongly from the system development directions suggested by automatic MT evaluation procedures. Current automatic procedures do not take into consideration factors such as the cognitive effort required for the resolution of different types of errors, or the emotional reactions they provoke in users. For example, errors that are inadvertently comical or culturally offensive might provoke strong negative user reactions and should thus be weighted more strongly by system developers when user acceptance is a key factor in the intended application. On the other hand, most users might expect – and thus be forgiving of – minor grammatical errors. A deeper insight into which errors are perceived as the most egregious for a particular MT application (depending on language pair, domain, etc.) is therefore crucial for improving user acceptance.

The latter point is exemplified by the project that provides the background for this study, viz. the TransPHorm project (Kirchhoff et al 2011). This project investigates the use of MT in the public health domain, with a focus on regional public health departments in the U.S. Pacific Northwest. One of the goals is to explore the feasibility of integrating MT into their standard workflow for producing multilingual health information materials. In this setting the potential users of MT are not trained translators or post-editors but bilingual public health professionals, i.e. they are domain experts but lay users with respect to MT technology. These individuals are unlikely to use MT unless a high level of user satisfaction can be achieved.

Despite a wealth of existing research on computational approaches to eliciting and modeling user preferences, little of it has been applied to MT evaluation. In this paper we explore the use of *conjoint analysis* (CA) to gain insights into users' preferences regarding different types of MT errors. CA is a formal framework for preference elicitation that was originally developed in mathematical psychology and is widely used in marketing research. Its typical application is to determine the reasons for consumers' purchasing choices. In CA studies, participants are presented with surveys asking them to choose from, rate, or rank a range of products characterized by different combinations of attributes. Statistical modeling, usually some form of multinomial regression analysis, is then used to infer the values ('utilities' or 'part-worths') consumers attach to different attributes. In a typical marketing setup the attributes might be price, packaging, performance, etc. In our case the attributes represent different types of MT errors and their frequencies. The outcome of CA is a list of values attached to different error types across a group of users, along with statistical significance values. CA survey outcomes often depend strongly on the quality of the surveys presented to respondents, and on the respondent population itself. Expanding on our earlier work (Kirchhoff et al 2012) we therefore utilize and compare two different ways of recruiting respondent populations and administering the survey. In the first case we use crowd-sourcing via the Amazon Mechanical Turk platform; that is, the respondents are not known to us and the quality of the responses might be lower. On the other hand, a large sample of responses can be gathered quickly. In the second case we present the survey to a sample of

domain experts who were recruited in person and tend to deliver high-quality data. However, the sample is of necessity much smaller under these conditions.

In the remainder of this paper we first review related background literature (Section 2) and give an overview of the basic techniques of CA (Section 3). This is followed by a description of the data set (Section 4) and experimental design (Section 5). Results and discussion are provided in Sections 6 and 7.

2 Background: MT Evaluation and User Preference Modeling

Current work in MT evaluation research falls into three main categories: *automatic evaluation*, *human evaluation*, and *embedded application evaluation*. Much effort has focused on the first category, i.e., on the problem of designing evaluation metrics that can be computed automatically for the purpose of system tuning and development. These include metrics such as BLEU (Papineni et al 2002), position-independent word error rate (PER), METEOR (Lavie and Agarwal 2007), and translation edit rate (TER) (Snover et al 2006). Given a set of references created by human translators, automatic evaluation metrics can be computed on-the-fly, without recourse to human evaluators for each new translation that is created. An alternative approach is to treat MT quality as a prediction task and to train a quality scorer on features extracted from the input and translations only (Specia et al 2012).

Human evaluation (see Denkowski and Lavie (2010) for a recent overview) typically involves rating translation output with respect to fluency and adequacy (LDC 2005). Alternatively, two or more translation outputs can be compared and ranked directly (Callison-Burch et al 2007). What is common to both automatic and human evaluation metrics is that they typically provide a *global* assessment of overall translation performance, i.e. a single score computed at the sentence, document, or corpus level. However, they do not provide insight into different *types* of MT errors. More fine-grained analyses of individual MT errors often include manual or (semi-)automatic error annotation to gain insights into the strengths and weaknesses of MT engines (Vilar et al 2006; Condon et al 2010; Popovic and Ney 2011; Farrús et al 2012).

Embedded application evaluation looks at how MT quality affects the performance of either human readers who use MT output for tasks such as question answering, gisting, etc., or automated back-end applications such as information retrieval. For example, several studies have looked at how MT errors influence the work of post-editors with respect to productivity and speed (Krings 2001; O'Brien 2011). In Al-Maskari and Sanderson (2006) it was analyzed how MT errors affected a multilingual Question Answering system. Yamashita and Ishida (2006) provided a study of MT errors in an embedded multilingual communication system. The impact of MT errors on information retrieval was investigated in Parton and McKeown (2010) and Parton et al (2012).

However, studies addressing the user acceptability of MT errors are few and mostly predate the recent success of state-of-the-art statistical MT engines. In Hui (2002) a heuristic, task-based evaluation was designed to identify users' responses to features of MT output such as conciseness and cohesiveness.

There is a wealth of research on user preference elicitation, much of which has been carried out in the context of developing decision support systems. Typical approaches include qualitative decision theory (Doyle and Thomason 1999), constraint satisfaction techniques (Boutilier et al 1997), and analytic hierarchy process (Saaty 1977). Overviews of different methods can be found in Chen and Pu (2004) and Braziunas (2006). The CA approach we use in our study is similar to classical decision theory. Our reason for choosing this approach is that it has found wide application in the social sciences and thus has been validated

extensively in practice; moreover, it consists of tasks that are easy to perform for respondents.

3 Conjoint Analysis

CA (Green and Rao 1971; Green and Srinivasan 1978) is based on discrete choice theory and studies how the characteristics of a product or service influence users' choices and preferences. It is typically used to evaluate and predict purchasing decisions in marketing research but has also been used in analyzing migration trends (Christiadi and Cushing 2007), decision-making in healthcare settings (Philips et al 2002), transportation choices (Maier and Edward 2002), and many other fields. The assumption is that a product or 'concept' can be described by a set of discrete attributes and their values or 'levels'. For example, a laptop can be described by CPU type, amount of RAM, price, battery life, etc. CA generates different concepts by systematically varying possible combinations of attributes and values and presents a subset of choices to respondents who then choose their preferred concept. The most and least preferred combinations are often obvious: for example, most respondents would choose a laptop with maximum CPU, RAM, and battery life for the minimum price. On the other hand, an expensive laptop with minimum CPU power, RAM, and battery life would be the most dispreferred option. The value of CA derives from studying combinations lying between these extremes since they shed light on the trade-offs customers are willing to make. For example, do customers tend to prefer CPU power over battery life, battery life over RAM, or some other constellation of attributes?

In an appropriately designed CA study, each attribute level is equally likely to occur. For a small number of attributes and levels, the total number of possible concepts (defined by different combinations of attributes) is generated and tested exhaustively; if the number of possible combinations is too large, sampling techniques are used. The total set of responses is then evaluated for main effects (i.e. the relative importance of each individual attribute) and for interactions between attributes.

Various different approaches to CA have been developed. The traditional full-profile CA requires respondents to rate or rank all concepts presented. In so-called 'choice-based conjoint analysis' (CBC) (Louviere and Woodworth 1983), several different concepts are presented, and respondents are required to choose one of them. Finally, adaptive CA dynamically adapts and changes the set of concepts presented to respondents based on their previous choices.

3.1 Choice-based conjoint analysis

CBC is currently the most widely used method of conjoint analysis, due to its simplicity: respondents merely need to choose one of a set of proposed concepts, a task which is similar to many real-life decision-making problems. The potential disadvantage is that the elicitation process is less efficient: respondents need to process the entirety of information presented before making a choice; therefore, it is advisable to only include a small number of concepts to choose from in any given task. CBC is thus appropriate for concepts involving a small number of attributes.

The most frequently used underlying statistical model for CBC is McFadden's conditional logit model (McFadden 1974). The conditional logit model specifies the n possible concept choices as a categorial dependent variable Y with outcomes 1, …, n. The decision of an individual respondent i in favor of the j'th outcome is based on a utility value u_{ij}, which must exceed the utility values for all other outcomes $k = 1, \ldots, n, k \neq j$. It is assumed that u_{ij} decomposes into a systematic or representative part υ_{ij} and a random part ε_{ij}; $u_{ij} = \upsilon_{ij} + \varepsilon_{ij}$. A further assumption is that the random components are independent and identically

distributed according to the extreme value distribution with the cumulative density function in (1):

$$F(\varepsilon_{ij}) = e^{-e^{-\varepsilon_{ij}}} \quad (1)$$

The systematic part v_{ij} is modeled as a linear combination $\beta' \mathbf{X}$, where $\mathbf{X} = \{x_1, \ldots, x_m\}$ is a vector of m observed predictor variables (the attributes of the alternatives), and β is a vector of coefficients indicating the importance of the attributes. Then, the probability that the *i'th* individual chooses the *j'th* outcome, $P(j|i)$, can be defined as in (2):

$$P(j|i) = \frac{e^{\beta' \mathbf{X}_{ij}}}{\sum_{k=1}^{n} e^{\beta' \mathbf{X}_{ik}}} \quad (2)$$

The β parameters are typically estimated by maximizing the conditional likelihood using the Newton-Raphson method (Ypma 1995). For basic CBC an aggregate logit model is used, where responses are pooled across respondents. In this case a single set of β parameters is used to represent the average preferences of the entire respondent population, rather than individuals' preferences. This implicitly assumes that respondents form a homogeneous group, which is often not correct. This oversimplification can be circumvented by applying latent class analysis (Goodman 1974), which groups respondents into homogeneous subsets and estimates different utility values for each one.

There are numerous advantages to using a formal analysis framework of this type rather than simply questioning users about their experience.

1. When assessing a complex 'product' like MT output, users are notoriously poor at analyzing their own judgments and stating them in explicit terms, especially when they lack linguistic training. It has been noted in the past that it is often difficult for human evaluators to assign consistent ratings for fluency and adequacy, leading to low inter-annotator agreement (Callison-Burch et al 2007). Requiring users to rank the output from different systems has proven easier but, as discussed in Denkowski and Lavie (2010), it is still difficult for evaluators to produce consistent rankings. By contrast, the CA framework used here only requires the choice of one out of several possibilities. Users are not asked to provide an objective ranking of several translation possibilities but a single, personal choice, which is an easier task. Furthermore, the choice-based design provides a way of observing trade-offs that users make with respect to different types and numbers of errors. For instance, from the user's point of view, do three morphological errors in one sentence count as much, more, or less than a single word-sense error?
2. CA provides numerical values ('utilities' or 'part-worths') indicating the relative importance of different features of a particular MT output. These might be helpful in tuning MT engines in the future, provided that different error types can be classified automatically.
3. It is also possible to analyze interactions between different attributes, e.g. the effect that a certain combination of errors (both word order and word sense error present in one sentence, say) has vs. other combinations.
4. Different techniques exist to segment the population into different user types (or 'market segments') and estimate different utility values for each. However, in this paper only aggregate CA will be used, where preferences are analyzed for the entire pool of respondents.

3.2 Conjoint analysis for eliciting machine translation user preferences

In our present study we view different MT outputs as different products or 'concepts' between which users may choose. We assume that users clearly prefer some machine translations over others, and that these preferences are dependent on the types and frequencies of the errors present in the translation. Thus, error types serve as the attributes of our concepts and the (discretized) error frequencies (e.g. high, medium, low) are the levels. Note that there may be other attributes of a translation (e.g. sentence length) that may affect a user's choice; these are not considered in this study but they could easily be included in future studies.

In contrast to most standard applications of CA, a particular combination of attributes defines not only a single concept but a large set of concepts (alternative translations of a single sentence, or multiple sentences). It is, therefore, necessary to consider a sample of sentences for each combination of attributes. Thus, compared to equation (2), we have another conditioning variable s ranging over sentences, as in (3):

$$P(j|i,s)=\frac{e^{\beta' \mathbf{x}_{ijs}}}{\sum_{k=1}^{n} e^{\beta' \mathbf{x}_{ijs}}} \quad (3)$$

Our procedure for this study is as follows. First, we select the error types to be investigated. This is done by manually annotating MT errors in our data set and selecting the most frequent error types. The error frequencies are quantized into a small number of levels for each error type. We then generate different profiles (combinations of attributes/levels) and group them into choice tasks: these are the combinations of profiles from which respondents will choose one. The respondents' choices are gathered and evaluated. A single set of model parameters is estimated, aggregating over both respondents and sentences, and statistical significance values are computed. Additionally, we perform prediction experiments, using the estimated utility values to predict users' choices on held-out data.

4 Data and Error Annotation

The corpus collected for this project consists of informational materials on general health and safety topics (e.g. HIV, STDs, vaccinations, emergency preparedness, maternal and child health, diabetes, etc.) collected from a variety of English-language public health websites. The documents were translated into Spanish by Google Translate.[1] 60 of these documents were then manually annotated for errors by two native speakers of Spanish. Our error annotation scheme is similar to other systems that have been developed for Spanish in the past (Vilar et al 2006) and comprises the following categories:

1. **Untranslated word**. These are original English words that have been left untranslated by the MT engine and that are not proper names or English words typically used in Spanish.
2. **Missing word.** A word necessary in the output is missing; a further distinction is made between missing function words and missing content words.
3. **Word sense error**. The translation reflects a word sense of the English word that is wrong or inappropriate in the present context.
4. **Morphology**. The morphological features of a word in the translation are wrong.

[1] http://translate.google.com

5. **Word order error**. The word order is wrong – a further distinction is made between short-range errors (within a linguistic phrase, e.g. adjective-noun ordering errors) and long-range errors (spanning a phrase boundary).
6. **Spelling**. Orthographic error.
7. **Superfluous word**. A word in the translation is redundant or superfluous.
8. **Diacritics**. The diacritics are faulty (missing, superfluous, or wrong).
9. **Punctuation**. Punctuation signs are missing, wrong, or superfluous.
10. **Capitalization**. Missing or superfluous capitalization.
11. **Pragmatic/Cultural error**. The translation is unacceptable for pragmatic or cultural reasons, e.g. offensive or comical.
12. **Other**. Anything not covered by the above categories.

Some of these categories (e.g. missing or superfluous word) are fairly general and apply to MT output regardless of the source and target language. Others, such as morphology and diacritics, are more specific to Spanish and would not be relevant to certain other languages (e.g. Chinese). Finally, the category of pragmatic/cultural errors was included because this was perceived to be an important aspect of evaluating translations of health information materials.

Annotators were linguistically trained and were supervised in their annotation efforts. Then, for a randomly selected subset of 25 of these documents (1804 sentences), they were instructed to create a consensus error annotation, and to subsequently correct the errors, thus producing consensus reference translations.

Computing BLEU/PER scores against the corrected output yields a BLEU score of 0.658 and a PER of 19.8%. Not surprisingly, these scores are very good since the reference translations are corrections of the original output rather than independently created translations. However, annotators independently judged the overall MT quality as quite good as well.

The detailed errors statistics computed from the 25 documents is shown in Table 1. The most frequent error types are, in order: morphological errors, word sense errors, missing function words, and word order errors.

5 Study Design

Based on the error statistics we defined four error types to be used as the attributes in our CA study: word sense errors (S), morphology errors (M), word order errors (O), and function word errors (F) (where the latter includes both missing and superfluous function words). For word sense, word order, and function word errors we defined two values (levels): high (H) and low (L). Since morphology errors are much more frequent than others, we use a three-valued attribute in this case (high, medium (M), and low).

We selected 40 different sentences from our annotated documents, with the constraint that each sentence had to contain a minimum of one instance each of sense, order and function word errors, and a minimum of two instances of morphological errors. Based on the error annotations and their manual corrections, each sentence was edited selectively to reflect different attribute levels, i.e. different numbers of errors of a given type. For example, different versions of a sentence can be created that exhibit a high, medium, or low level of morphological errors. The variable numbers of errors were mapped to the discrete attribute

levels as follows: if the total number of errors for a given type is ≤ 2, then H equals 2 errors and L equals 0 errors for the binary attributes, and H=2, M=1, L=0 for the three-valued attribute. When the number of errors is larger than 2, the interval size for each level is defined by the number of errors divided by the number of levels, rounded to the nearest integer. Editing is done semi-automatically: first, simple modifications (e.g. word substitutions) are performed automatically based on an alignment of the machine translation and its manually corrected version. Second, manual editing is performed to address more complex errors (e.g. word order errors) and to ensure the overall correctness of the modified sentence.

The number of all possible different combinations of attributes/levels is 24; thus, for each sentence, 24 concepts or "profiles" are constructed. A partial example is shown in Table 2.

6 Experiments and Results

We chose a full factorial experiment design, i.e. each of the 24 possible profiles was utilized for each of the 40 sentences. Each partially edited sentence represents a different profile. However, not all 24 profiles can be presented simultaneously to a single respondent; typically, CBC surveys need to be kept as small and simple as possible to prevent respondents from resorting to simplification strategies and delivering noisy response data. Profiles were grouped into choice tasks with three alternatives each, representing a balanced distribution of attribute levels. For each survey, four choice tasks were randomly selected from the total set of choice tasks. The questions in the survey thus included profiles pertaining to different sentences, which was intended to avoid respondent fatigue.

6.1 Recruitment strategies

This set of surveys was used to conduct two separate CA studies, distinguished by their participant pool and recruitment method. In the first study, participants were recruited through Amazon Mechanical Turk,[2] an online platform for micro-task brokering. Participants who have a Mechanical Turk account (and may in principle be located all over the world) can accept so-called Human Intelligence Tasks (HITs) placed online by requesters; these are small tasks that are easy to do for a human but difficult to do for a computer. Each task is paid at a very low rate; due to the large number of workers and their 24/7 availability, responses can be collected rapidly and at a low cost. The drawback is that workers are not known to requesters, and certain steps need to be taken to ensure quality control and prevent workers from delivering low-quality output. The surveys described above were published on Mechanical Turk and were restricted to workers who had previously delivered high-quality results on other Spanish translation and annotation HITs we had published on Mechanical Turk. For each choice task, workers were instructed to carefully read the original source sentence and the translations provided, then choose the one they liked best (an obligatory choice question with the possibility of choosing exactly one of the alternatives provided), and to state the reason for their preference (an obligatory free-text answer); the latter was included as a quality control step to prevent workers from making random choices. In total we published 240 HITs (surveys) with four choice tasks and three assignments each, resulting in a total of 2880 responses. A total of 29 workers completed the HITs, with a variable number of HITs per worker. We did not collect any demographic information about workers.

In the second study we recruited a set of 15 bilingual staff members from the Washington State Department of Health who had previously collaborated with us on our project. Eight of

[2] http://www.mturk.com

them spoke mostly Mexican Spanish, six participants represented a variety of Latin American dialects (Salvadoran, Venezuelan, etc.), and one participant spoke Castilian Spanish. All of them received a survey containing 34 choice tasks (selected from the 40 tasks used previously) with three choices each, resulting in a total of 510 responses. In contrast to the Mechanical Turk experiment, the population here consists of domain experts, all of whom had worked in public health for a number of years. Due to the limited availability of domain experts, the participant pool was smaller than in the crowd-sourced study, and each participant had to answer a larger number of questions in order to obtain the desired minimum number of responses overall. In the crowd-sourced study, workers were free to accept only as many HITs as they felt comfortable completing. Furthermore, all participants in the domain expert sample answered the same set of questions, whereas different respondents answered different though overlapping sets of questions in the crowd-sourced experiment, since it is more difficult to control how many and which HITs a worker accepts. The instructions issued to the domain expert sample were identical to those used in the crowd-sourced study except for the request to indicate a reason for their response; since the participants were known to us from a previous study we felt that a sanity check of the type used in the Mechanical Turk setup was not necessary. Both studies were approved by the University of Washington Institutional Review Board.

6.2 Quantitative analysis

We first measured the overall agreement among the three different responses per choice task using Fleiss' Kappa (Fleiss 1971). Fleiss' Kappa is a measure of the agreement among a fixed number of respondents giving categorical ratings, corrected for the level of chance agreement. It is computed as in (4):

$$K = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \quad (4)$$

where $\bar{P}$ is the average level of agreement and $\bar{P}_e$ is the chance level of agreement. K thus indicates the actually achieved fraction of the maximum above-chance level of agreement. The Kappa coefficient was 0.35 for the crowd-sourced respondents, and 0.30 for the public health professionals. The lower agreement in the latter group is not surprising since it was measured across 15 respondents versus three respondents per task in the former group. Fleiss' Kappas between 0.21 and 0.40 are genally considered 'fair' agreement (Landis and Koch 1977; Altman 1991), but our values indicate that there is still a substantial amount of variation among subjects regarding their preferred translation choice.

We next estimated the coefficients of the conditional logit model, first concentrating on the main effects only. We used the conditional logit model implementation in the R package.[3] The model's β coefficients, exponentiated β's, and significance values computed from the crowd-sourced responses are shown in Table 3. It is easiest to interpret the exponentiated β coefficients: these represent the change in the odds (i.e. odds ratios) of the error type being associated with the chosen translation, for each unit increase in the error level and while holding other error levels constant. For example, if the level of word sense errors is increased by 1 (i.e. it goes from low to high) while other error types are being held constant, the odds of the corresponding translation being chosen decrease by a multiplicative factor of 0.5325 (i.e. roughly 50%). Overall we see that word order errors are the most dispreferred, followed by word sense, morphology, and function word errors. All values are highly significant ($p < 0.001$, two-sided z-test).

[3] http://www.r-project.org

The corresponding analysis for the domain expert sample is shown in Table 4. Here, we see the same ranking of error types: word order errors are the most disliked error type, followed by word sense, morphology, and function words. The first two are again highly significant ($p < 0.001$), morphology error coefficients are significant ($p < 0.01$), and function word errors are not significant in this sample. Note that the actual coefficients are not directly comparable to those in Table 3 since a much smaller sample size (510 vs. 2880) was used in this case. Nevertheless, the preference relationships are the same.

We next tested all pairwise interactions between individual attributes. An interaction between two attributes means that the impact of one attribute on the outcome is dependent on the level of the other attribute. In the crowd-sourced sample we found two statistically significant interactions, between word sense and function word errors, and between morphological and function word errors, shown in Table 5. The meaning of the coefficients in Table 5 changes with the introduction of interaction terms, and they cannot directly be compared to those in Table 3. In particular, the $exp(\beta)$ for M:F and S:F now need to be interpreted as *ratios of odds ratios* for unit increases in the attribute levels. The values (> 1) indicate that the odds ratio of a positive choice associated with a unit increase in function word error level actually increases as the level of M or S errors rises, e.g. the odds ratio for S=*high* is 0.4462 ($exp(\beta_S + \beta_{S:F})$ vs. 0.3398 for S=*low*). This means that function word errors have a stronger impact on respondents' choices at low levels of morphological or word sense errors; by contrast, when the level of morphological/word sense errors is high, respondents are less sensitive to function word errors. This effect is also observable for word order and function word errors, but it is not statistically significant.

In the domain expert sample we saw a weak compounding effect between word sense and morphological errors, suggesting that respondents' likelihood of choosing a translation drops rapidly when the levels of both of these error types are high. However, this effect was barely statistically significant ($p < 0.05$). No other significant interactions were found.

A standard way of validating the overall explanatory power of the model is to perform prediction on a held-out data set. To this end we compute the probability of each choice in a set according to equation (3) by inserting the estimated β coefficients and take the max over *j*, which can be simplified as (5):

$$j^* = max_j \, \beta' X_{ijs} \quad (5)$$

The percentage of correctly identified outcomes (the "hit rate" or accuracy) is then used to assess the quality of the model. We perform testing by n-fold cross-validation (n=8 for the crowd-sourced sample and n=5 for the smaller domain expert sample). In each rotation one fold of the samples pertaining to each sentence is assigned to the test set; the rest are assigned to the training set. The coefficients of the conditional logit model are trained on the training portion and are then used to predict the outcomes on the test portion according to equation (5). The results from all folds were averaged and the standard deviation was computed. We compare the prediction results to two baseline models. The first is the random baseline: each training/test sample is a choice task with three alternatives; thus, choosing one alternative randomly results in a baseline accuracy of 33.33%. The second baseline consists of choosing the translation with the lowest number of errors overall.

Table 6 shows the results. On the crowd-sourced sample, the fewest-errors model obtains accuracies between 45.75% and 53.75%, with an average of 49.59%. The accuracies obtained by our model with the fitted coefficients range from 53.00% to 58.75%, with an

average of 54.06%. This is significantly better than the baseline models ($p < 0.001$, difference of proportions significance test).

The results for the domain experts are slightly different. The fewest errors baseline obtained accuracies between 37.25% and 41.18%, with an average of 38.40%. The conditional logit model achieved accuracies ranging from 47.05% to 54.90%, with an average of 51.21%. Thus, the conditional logit model significantly outperforms both baselines while the fewest-errors baseline outperforms the random baseline but not by a significant margin.

Finally, we also analyzed the free-text answers about the reason for workers' choices that were elicited in the Mechanical Turk experiment. They generally fall into two categories: Some answers were fairly generic ("It was the best translation", "the translation with the fewest errors", "it was the easiest to correct"), while others highlighted specific error types or preferred translations of particular words or phrases ("this option uses the best word order", "it has the best translation of *significant lead exposure*"). However, a clear pattern of error type preferences could not be established from the free-text answers, since workers were too inconsistent in the level of detail they provided in their answers.

7 Discussion

The main results from our experiments look surprisingly consistent, despite the different recruitment methods, subject populations, and sample sizes involved. Word order errors are the most dispreferred error type, followed by word sense, morphology, and function word errors. Statistically significant interactions occured in the crowd-sourced sample; they indicated that function word errors seem to be masked by more "serious" errors like word sense and word order errors when the level of those errors is high. When their level is low, function word errors tend to have a greater influence on respondents' choices. In the experts' sample only a weakly significant interaction between word sense and morphology errors was found: at high levels of word sense errors, additional morphology errors further decrease the likelihood of the translation being chosen. Thus, interactions between error types may be of a masking as well as a compounding type. This may be related to the experiment design and the respondent population. In a fast-paced setting such as the Mechanical Turk environment, workers tend to progress rapidly through a large number of HITs. In our setup, workers had to read three similar translations carefully and compare them in order to find all errors. Small errors like function word errors may go unnoticed; indeed, some workers submitted comments claiming that two translations were identical when in fact they were distinguished by one or two function word errors. On the other hand, the subjects in the domain expert sample generally pursue a very high standard of quality in their work and have been trained to be very accurate in producing and translating health and safety communication materials. Informal feedback from the domain experts obtained after the survey was completed indicates that they approached the survey with a high degree of diligence, and that they spent more time than expected on the survey. Thus, they may have been less prone to 'blending out' the less obvious error types.

The difference between the two samples is also evident in the prediction accuracy rates obtained by our model vs. the fewest-errors baseline. In both cases, the conditional logit model achieves the highest accuracy, but the difference between them is much more pronounced. The crowdworkers largely – though not exclusively – choose translations with an overall low number of errors. For domain experts this is clearly not the preferred strategy; differential modeling of error types leads to a jump in prediction accuracy. It is likely that the domain experts intuitively evaluated the different types of errors with respect to possible uses of MT in their own work, although they were not explicitly instructed to do so.

Clearly there is room for improvement in the predictive accuracy of the model. The model shows similar prediction rates on the training data; thus, the model's ability to generalize to unseen test data is not the problem here. Rather, the difficulty lies in the underlying variability of the data to be modelled, in particular the diversity of the user group and the sentence materials. For example, no distinction has been made between short-range and long-range word order errors, although it may be assumed that long-range word order errors are considered more severe by users than short-range errors. Another source of variability is the respondent population itself; since we only used aggregate conjoint analysis in this study, preferences are averaged over the entire population, ignoring potential sub-types of users. It may well be possible that some user types are more accepting of (say) word-order errors than word sense errors, or vice versa; recall that the agreement coefficients on the choices were only 0.30 and 0.35, respectively.

8 Conclusions and Future Work

We have studied the use of CA as a formal framework for eliciting user preferences for different types of MT errors. Our results confirm that, at least for the language pair and populations studied, users do not necessarily rely on the overall number of errors when expressing their preferences for different MT outputs. Instead, some error types affect users' choices more strongly than others. Of the different error types considered in this study, word order errors have the lowest frequency in our data but are the most dispreferred error type, followed by word sense errors. The most frequent error type in our data, namely morphology errors, is ranked third, and function word errors are the most tolerable. The viability of the CA framework was demonstrated by showing that the prediction accuracy of the fitted model exceeds that of a random or fewest-errors baseline.

In future work the overall predictive power of the model could be improved by more fine-grained modeling of different sources of variability in the data. Latent class analysis will be used in order to obtain preference models for different user types. In the long run, such models could be exploited for rapid user adaptation of MT engines after eliciting a few basic preferences from the user. Utility values obtained by conjoint analysis might also be used in MT system tuning, by appropriately weighting different error types in proportion to their utility values; however, this would require high-accuracy automatic classification of different error types.

Another way of extending the present analysis is to elicit user preferences in the context of a specific task to be accomplished; for instance, users could be asked to indicate their preferred translation when faced with the tasks of post-editing or extracting information from the translation. Finally, it is also possible to investigate a larger set of error types than those considered in this study. These may include different types of word order errors (long-range vs. short-range), consistency errors (where a source term is not translated consistently in the target language throughout a document), or named-entity errors.

Acknowledgments

We are grateful to Aurora Salvador Sanchis and Lorena Ruiz Marcos for providing the error annotations and corrections, to Megumu Brownstein for recruiting the domain experts, and to Kate Cole for comments on an earlier draft of this paper. This study was funded by grant #1R01LM010811-01 from the National Library of Medicine (NLM). Its content is solely the responsibility of the authors and does not necessarily represent the view of the NLM.

References

Al-Maskari, A.; Sanderson, M. The affect [sic] of machine translation on the performance of Arabic-English QA system. EACL-2006, 11th Conference of the European Chapter of the Association for

Computational Linguistics, Proceedings of the workshop on Multilingual Question Answering MLQA06; Trento, Italy. 2006. p. 9-14.

Altman, D. Practical Statistics for Medical Research. Chapman & Hall; London: 1991.

Boutilier, C.; Brafman, R.; Geib, C.; Poole, D. A constraint-based approach to preference elicitation and decision making. AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning; Stanford, CA. 1997. p. 19-28.

Braziunas, D. Tech rep. Department of Computer Science, University of Toronto; Canada: 2006. Computational approaches to preference elicitation.

Callison-Burch, C.; Fordyce, C.; Koehn, P.; Monz, C.; Schroeder, J. (Meta-)evaluation of machine translation. ACL 2007: Proceedings of the Second Workshop on Statistical Machine Translation; Prague, Czech Republic. 2007. p. 136-158.

Chen, L.; Pu, P. Tech Rep IC/2004/67. Human Computer Interaction Group, Ecole Politechnique Fédérale de Lausanne; Switzerland: 2004. Survey of preference elicitation methods.

Christiadi, Cushing B. Conditional logit, IIA, and alternatives for estimating models of interstate migration. 46th Annual Meeting of the Southern Regional Science Association; Charleston, SC. 2007. available online at http://rri.wvu.edu/wp-content/uploads/2012/11/wpcushing2007-4.pdf

Condon, S.; Parvaz, D.; Aberdeen, J.; Doran, C.; Freeman, A.; Awad, M. Evaluation of machine translation errors in English and Iraqi Arabic. LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation; Valetta, Malta. 2010. p. 729-735.

Denkowski, M.; Lavie, A. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. AMTA 2010: Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas; Denver, CO, USA. 2010. available online at http://amta2010.amtaweb.org/

Doyle J, Thomason R. Background to qualitative decision theory. AI Magazine. 1999; 20(2):55–68.

Farrús M, Costa-Jussà M, Popovic M. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. Journal of the American Society for Information Science and Technology. 2012; 63(1):174–184.

Fleiss J. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76(5): 378–382.

Goodman L. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974; 61(2):215–231.

Green P, Rao V. Conjoint measurement for quantifying judgmental data. Journal of Marketing Research. 1971; 8(3):355–363.

Green P, Srinivasan V. Conjoint analysis in consumer research: Issues and outlook. Journal of Consumer Research. 1978; 5:103–123.

Hui, B. Measuring user acceptability of machine translations to diagnose system errors: An experience report. Coling-2002 workshop “Machine translation in Asia”; Taipei, Taiwan. 2002. p. 63-70.

Kirchhoff K, Turner A, Axelrod A, Saavedra F. Application of statistical machine translation to public health information: a feasibility study. Journal of the American Medical Informatics Association. 2011; 18:472–482.

Kirchhoff, K.; Capurro, D.; Turner, A. Evaluating user preferences in machine translation using conjoint analysis. EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation; Trento, Italy. 2012. p. 119-126.

Krings, H. Empirical Investigations of Machine Translation Post-Editing Processes. Kent State University Press; Kent, OH: 2001.

Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159174.

Lavie, A.; Agarwal, A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation; Prague, Czech Republic. 2007. p. 228-231.

LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Linguistic Data Consortium; Philadelphia, PA: 2005. revision 1.5. Tech. rep.

Louviere J, Woodworth G. Design and analysis of simulated consumer choice experiments: an approach based on aggregate data. Journal of Marketing Research. 1983; 20(4):350–67.

Maier G, Edward M. Modelling preferences and stability among transport alternatives. Transportation Research Part E. 2002; 38:319–334.

McFadden, D. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P., editor. Frontiers in Econometrics. Academic Press; New York: 1974. p. 105-142.

O'Brien, S., editor. Cognitive Explorations of Translation: Eyes, Keys, Taps. Continuum; London/ New York: 2011.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, WJ. BLEU: a method for automatic evaluation of machine translation. 40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference; Philadelphia, PA, USA. 2002. p. 311-318.

Parton, K.; McKeown, K. MT error detection for cross-lingual question answering. Coling 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference; Beijing, China. 2010. p. 946-954.

Parton, K.; Habash, N.; McKeown, K.; Iglesias, G.; Gispert, A. Can automatic post-editing make MT more meaningful?. Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT); Trento, Italy. 2012. p. 111-118.

Philips K, Maddala T, Johnson F. Measuring preferences for health care interventions using conjoint analysis. Health Services Research. 2002; 37(6):1681–1705. [PubMed: 12546292]

Popovic M, Ney H. Towards automatic error analysis of machine translation output. Computational Linguistics. 2011; 37(4):657–688.

Saaty T. A scaling method for priorities in hierarchical structure. Journal of Mathematical Psychology. 1977; 15:234–281.

Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Visions for the Future of Machine Translation; Cambridge, MA, USA. 2006. p. 223-231.

Specia L, Raj D, Turchi M. Machine translation evaluation versus quality estimation. Machine Translation. 2012; 24(1):39–50.

Vilar, D.; Xiu, J.; D'Haro, L.; Ney, H. Error analysis of statistical machine translation output. LREC-2006: Fifth International Conference on Language Resources and Evaluation, Proceedings; Genoa, Italy. 2006. p. 697-702.

Yamashita, N.; Ishida, T. Effects of machine translation on collaborative work. Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW); Banff, CA. 2006. p. 515-523.

Ypma T. Historical development of the Newton-Raphson method. SIAM Review. 1995; 37(4):531–551.

Table 1

Error statistics from manual consensus annotation of 25 documents. The two right-hand columns show error subtypes.

Type	%	Subtypes	%
Morphology	28.2	Verbal	15.8
		Nominal	12.4
Missing word	16.7	Function word	12.6
		Content word	4.1
Word sense error	16.1		
Word order error	9.7	Short range	8.0
		Long range	1.7
Punctuation	9.1		
Other	5.9		
Spelling	5.1		
Superfluous word	4.7	Function word	3.8
		Content word	0.9
Capitalization	2.7		
Untranslated word	1.1	Medical term	0.0
		Proper name	0.2
		Other	0.9
Pragmatic	1.0		
Diacritics	0.2		
Total	100.0		

Table 2

Examples of the 24 attribute combinations and corresponding partially edited translations for the English input sentence *Planning ahead and taking a few short cuts will save both your time and your food dollars.*

No.	Attributes	Sentence
1	S=H:M=H:O=H:F=H	Planear con anticipación y tomar un atajo pocos ahorrar su tiempo y su dinero para alimentos.
2	S=H:M=H:O=H:F=L	Planear con anticipación y tomar un atajo le pocos ahorrar su tiempo y su dinero para la alimentos.
3	S=H:M=H:O=L:F=H	Planear con anticipación y tomar un pocos atajo ahorrar su tiempo y su dinero para alimentos.
4	S=H:M=H:O=L:F=L	Planear con anticipación y tomar un pocos atajo le ahorrar su tiempo y su dinero para la alimentos.
5	S=H:M=M:O=H:F=H	Planear con anticipación y tomar un atajo pocos ahorrar su tiempo y su dinero para alimentos.
6	S=H:M=M:O=H:F=L	Planear con anticipación y tomar un atajo le pocos ahorrar su tiempo y su dinero para la alimentos.
7	S=H:M=M:O=L:F=H	Planear con anticipación y tomar un pocos atajo ahorrará su tiempo y su dinero para alimentos.
8	S=H:M=M:O=L:F=L	Planear con anticipación y tomar un pocos atajo le ahorrará su tiempo y su dinero para la alimentos.
9	S=H:M=L:O=H:F=H	Planear con anticipación y tomar unos atajos pocos ahorrará su tiempo y su dinero para alimentos.
10	S=H:M=L:O=H:F=L	Planear con anticipación y tomar unos atajos le pocos ahorrará su tiempo y su dinero para la alimentos.
etc.	etc.	etc.
24	S=L:M=L:O=L:F=L	Planear con anticipación y realizar unos pocos recortes le ahorrará su tiempo y su dinero para la comida.

Table 3

Estimated coefficients in the conditional logit model and associated significance levels (α): main effects, based on crowd-sourced responses.

Variable	β	exp(β)	α
Word order	-1.125	0.3246	0.001
Word sense	-0.6302	0.5325	0.001
Morphology	-0.4034	0.6680	0.001
Function words	-0.1211	0.8859	0.001

Table 4

Estimated coefficients in the conditional logit model and associated significance levels (α): main effects, based on domain expert sample.

Variable	β	exp(β)	α
Word order	-0.8006	0.4491	0.001
Word sense	-0.5499	0.5770	0.001
Morphology	-0.2069	0.8131	0.010
Function words	-0.0479	0.9532	1.000

Table 5

Estimated coefficients in the conditional logit model and associated significance values (α): crowd-sourced sample, with interactions. Variables containing ":" denote interaction terms.

Variable	β	exp(β)	α
Word order (O)	-1.149	0.3169	0.001
Word sense (S)	-1.079	0.3398	0.001
Morphology (M)	-0.6971	0.4980	0.001
Function words (F)	-0.8932	0.4094	0.001
M:F	0.2081	1.231	0.001
S:F	0.2649	1.303	0.010

Table 6

Average cross-validation accuracies and standard deviations of conditional logit model, fewest-errors-baseline, and random baseline.

	Crowdworkers		Domain experts	
	Accuracy (%)	**Stddev**	**Accuracy (%)**	**Stddev**
Random	33.33	0.0	33.33	0.0
Fewest errors	49.49	2.70	38.40	1.90
Clogit	54.68	1.99	51.21	2.51