Evaluating User Preferences in Machine Translation Using Conjoint Analysis

Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
Seattle, WA, USA
katrin@ee.washington.edu

Daniel Capurro, Anne Turner
Department of Medical Education
and Biomedical Informatics
University of Washington
Seattle, WA, USA
dcapurro@u.washington.edu
amturner@u.washington.edu

Abstract

In spite of much ongoing research on machine translation evaluation there is little quantitative work that directly measures users' intuitive or emotional preferences regarding different types of machine translation errors. However, the elicitation and modeling of user preferences is an important prerequisite for future research on user adaptation and customization of machine translation engines. In this paper we explore the use of conjoint analysis as a formal quantitative framework to gain insight into users' relative preferences for different translation error types. Using English-Spanish as the translation direction we conduct a crowd-sourced conjoint analysis study and obtain utility values for individual error types. Our results indicate that word order errors are clearly the most dispreferred error type, followed by word sense, morphological, and function word errors.

1 Introduction

Current work in machine translation (MT) evaluation research falls into three different categories: *automatic evaluation*, *human evaluation*, and *embedded application evaluation*. Much effort has focused on the first category, i.e. on designing evaluation metrics that can be computed automatically for the purpose of system tuning and development. These include e.g. BLEU (Papineni et al., 2002), position-independent word error rate (PER), METEOR (Lavie and Agarwal, 2007), or translation error rate (TER) (Snover et al., 2006). Human evaluation (see (Denkowskie and Lavie, 2010) for a recent overview) typically involves rating translation output with respect to fluency and adequacy (LDC, 2005), or directly comparing and ranking two or more translation outputs (Callison-Burch et al., 2007). All of these evaluation techniques provide a *global* assessment of overall translation performance without regard to different error types.

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; Popovic and Ney, 2011; Condon et al., 2010; Farreus et al., 2012). There have also been studies of how MT errors influence the work of post-editors with respect to productivity, speed, etc. (Krings, 2001; O'Brien, 2011) or the performance of back-end applications like information retrieval (Parton and McKeown, 2010).

In contrast to this line of research, there is surprisingly little work that directly investigates which types of errors are intuitively the most disliked by users of machine translation. Although there is ample anecdotal evidence of users' reactions to machine translation, 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. User preferences might sometimes diverge strongly from the system development directions suggested by automatic evaluation procedures. Most 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 reac-

tions 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 machine translation application (depending on language pair, domain, etc.) is therefore crucial for improving user acceptance. In addition, user adaptation and customization of MT engines are emerging as important future directions for machine translation research, and it is necessary to develop principled strategies for eliciting and modeling user preferences. However, despite a wealth of existing research on computational preference elicitation techniques little of it has been applied to machine translation evaluation research.

In this paper we explore the use of *conjoint analysis* (CA) to gain knowledge of users' preferences regarding different types of machine translation errors. Conjoint analysis is a formal framework for preference elicitation that was originally developed in mathematical psychology and is widely used in marketing research (Green and Srinivasan, 1978). Its typical application is to determine the reasons for consumers' purchasing choices. In conjoint analysis studies, participants are asked to choose from, rate, or rank a range of products characterized by different combinations of attributes. Statistical modeling, typically 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 machine translation errors and their frequencies. The outcome of conjoint analysis is a list of values attached to different error types across a group of users, along with statistical significance values.

In the remainder of this paper we will first give an overview of the basic techniques of conjoint analysis (Section 2), followed by a description of the data set (Section 3) and experimental design (Section 4). Results and discussion are provided in Section 5. Section 6 concludes.

2 Conjoint Analysis

Conjoint analysis 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), and many other fields. The assumption is that each 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 the combination of attributes and values and letting respondents choose their preferred one. Clearly, the most preferred and least preferred combinations are known (e.g. a laptop with maximum CPU power, RAM and battery life at the minimum price would be the most preferred). The value of CA derives from studying intermediate combinations between these extremes since they shed light on the trade-offs users are willing to make. 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 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 conjoint analysis dynamically adapts and changes the set of concepts presented to respondents based on their previous choices. 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, as task which is similar to many real-life decision-making problems. The 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 con-

cepts 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 $k = 1, ..., n, k \neq j$. It is assumed that u_{ij} decomposes into a systematic or representative part v_{ij} and a random part ε_{ij}; $u_{ij} = v_{ij} + \varepsilon_{ij}$. A further assumption is that the random components are independent and identically distributed according to the extreme value distribution with cumulative density function

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

The systematic part v_{ij} is modeled as a linear combination $\beta'\mathbf{X}$, where $\mathbf{X} = \{x_1, ..., x_m\}$ 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:

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

The β parameters are typically estimated by maximizing the conditional likelihood using the Newton-Raphson method. 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 an entire market, rather than individuals' preferences. This implicitly assumes that respondents form a homogeneous group, which is typically 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. First, for a complex "product" like machine translation 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 (Denkowskie 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 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? Second, CA provides numerical values ("utilities" or "part-worths") indicating the relative importance of different features of a machine translation output. These might be helpful in machine translation system tuning provided that different error types can be classified automatically. Third, it is also possible to analyze interactions between different attributes, e.g. the effect that a certain combination of errors (e.g. both word order and word sense error present in one sentence) has vs. other combinations. Fourth, 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 conjoint analysis will be used, where preferences are analyzed for the entire population surveyed.

2.1 Conjoint analysis for eliciting machine translation user preferences

When applying the conjoint analysis framework to machine translation evaluation we treat different machine translations 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 features 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 conjoint analyis 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 useful to consider a representative sample of sentences for each combination of attributes. Thus, compared Eq. 2 we have another conditioning variable s ranging over sentences:

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

Our procedure for this study is as follows. First, we select the error types to be investigated. This is done by manually annotating machine translation errors in our data set and selecting the most frequent error types. The different 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. Respondents' choices are gathered through Mechanical Turk. Finally, we estimate a single set of model parameters, aggregating over both respondents and sentences, and compute statistical significance values. Additionally, we perform prediction experiments, using the estimated utility values to predict users' choices on held-out data.

3 Data

The data used for the present study was collected as part of a research project on applying machine translation to the public health domain. It consists of information 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 (http://www.google.com/translate). 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 used for Spanish (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 in use in Spanish.

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

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.
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.

Annotators were linguistically trained and were supervised in their annotation efforts.

For a subset of 25 of these documents (1804 sentences), the annotators 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 65.8 and a PER of 19.8%. Unsurprisingly, 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 translation 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. Based on this we defined four error types to be used as the attributes in our conjoint analysis study: word sense errors (S), morphology errors (M), word order errors (O) and function word errors (F) – 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).

From these documents we selected 40 sentences, each of which contained 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 can be edited selectively to reflect different attribute levels, i.e. different numbers of errors of a given type. For example, different versions of a sentence are created that exhibit a high, medium, or low level of morphological errors. The variable number of errors are mapped to the discrete attribute levels as follows: If the total number of errors for a given type is $\leq$ 2, then H = 2 errors and L = 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.

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.

4 Experiments

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, 4 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. Surveys were presented to respondents on the Amazon Mechanical Turk platform. For each choice task, Turkers 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 Turkers from making random choices. The set of Turkers was limited to those who had previously delivered high-quality results in other Spanish translation and annotation HITs we had published on Mechanical Turk. In total we published 240 HITs (surveys) with 4 choice tasks and 3 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. The responses were analyzed using the conditional logit model implementation in the R package.[1]

5 Results and Discussion

We first measured the overall agreement among the three different responses per choice task using Fleiss's Kappa (Fleiss, 1971). The kappa coefficient was 0.35, which according to (Landis and Koch, 1977) constitutes "fair agreement" but does indicate that there is considerable variation among workers regarding their preferred translation choice. We next estimated the coefficients of the conditional logit model considering main ef-

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

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.
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 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.*

.

Variable	β	exp(β)	α
O	-1.125	0.3246	0.001
S	-0.6302	0.5325	0.001
M	-0.4034	0.6680	0.001
F	-0.1211	0.8859	0.001

Table 3: Estimated coefficients in the conditional logit model and associated significance levels (α) – main effects. O = word order, S = word sense, M = morphology, F = function words.

Variable	β	exp(β)	α
O	-1.149e+00	3.169e-01	0.001
S	-1.079e+00	3.398e-01	0.001
M	-6.971e-01	4.980e-01	0.001
F	-8.932e-01	4.094e-01	0.001
M:F	2.081e-01	1.231e+00	0.001
S:F	2.649e-01	1.303e+00	0.01

Table 4: Estimated coefficients in the conditional logit model and associated significance values (α) – interactions. O = word order, S = word sense, M = morphology, F = function words. Variables containing ":" denote interaction terms.

fects only. The model's β coefficients, exponentiated β's, and significance values 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. 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). 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. We found two statistically significant interactions, between word sense and function word errors, and between morphological and function word errors. The meaning of the coefficients in Table 4 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 the latter 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.

	Accuracy (%)	Stddev
Clogit	54.68	1.99
Fewest errors	49.49	2.70
Random	33.33	0.0

Table 5: Average cross-validation accuracy and standard deviation of conditional logit model, fewest-errors-baseline, and random baseline.

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 Eq. 3 by inserting the estimated β coefficients and take the max over j, which can be simplified as:

$$j^* = max_j \beta' X_{ijs} \quad (4)$$

$$(5)$$

The percentage of correctly identified outcomes (the "hit rate" or accuracy) is then used to assess the quality of the model.

We performed 8-fold cross-validation. For each fold one eighth of the data for each sentence was assigned to the test set; the rest was assigned to the training set. Table 5 shows the average accuracies for our conditional logit model as well as two baselines. The first is the random baseline – each training/test sample is a choice task with 3 alternatives; thus, choosing one alternative randomly results in a baseline accuracy of 33.3%. The second baseline consists of choosing the alternative with the lowest number of errors overall. This leads to accuracies ranging from 45.75%-53.75%, with an average of 49.59%. The accuracies obtained by our model with the fitted coefficients range from 53.00%-58.75%, with an average of 54.06%. This is significantly better than the random baseline and clearly better (though not statistically significant) than the fewest-errors baseline. Nevertheless there clearly is room for improvement in the predictive accuracy of the model. The model shows virtually the same performance (54.04% accuracy on average) on the training data; thus, generalization ability 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-classes of users. It may well be possible that some user types are more accepting of e.g. word-order errors than word sense errors, or vice versa – recall that the agreement coefficient on the top choice was only 0.35. Finally, another confounding factor might be the quality of the Mechanical Turk data. Although we took several steps to ensure reasonable results, responses may not be as reliable as in a face-to-face study with respondents.

6 Conclusions and Future Work

We have studied the use of conjoint analysis to elicit user preferences for different types of machine translation errors. Our results confirms that, at least for the language pair and population studied, users do not necessarily rely on the overall number of errors when expressing their preferences for different machine translation 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, morphology errors, is ranked third, and function word errors are the most tolerable. The viability of the conjoint analysis 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. Specifically, we plan to compare the present results to results from face-to-face experiments, in order to gauge the reliability of crowd-sourced data for conjoint analysis. In addition, 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 machine translation 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 postediting 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. This study was funded by NLM grant 1R01LM010811-01.

References

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-)evaluation of machine translation. In *Proceedings of WMT*, pages 136–158.

Christiadi and B. Cushing. 2007. Conditional logit, IIA, and alternatives for estimating models of interstate migration. In *Proceedings of the 46th Annual Meeting of the Southern Regional Science Association*.

Condon, S., D. Parvaz, J. Aberdeen, C. Doran, A. Freeman, and M. Awad. 2010. Evaluation of machine translation errors in English and Iraqi Arabic. In *Proceedings of LREC*.

Denkowskie, M. and A. Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of AMTA*.

Farreus, M., M.R. Cosa-Jussa, and M. Popovic Morse. 2012. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. *Journal of the American Society for Information Science and Technology*, 63(1):174–184.

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

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

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

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

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

Lavie, A. and A. Agarwal. 2007. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 28–231.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. revision 1.5. Technical report, LDC.

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

McFadden, D.L. 1974. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press: New York.

O'Brien, S., editor. 2011. *Cognitive Explorations of Translation: Eyes, Keys, Taps*. Continuum.

Papineni, K., S. Roukos, and T. Ward. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Parton, K. and K. McKeown. 2010. MT error detection for cross-lingual question answering. In *Proceedings of Coling*.

Philips, K., T. Maddala, and F.R. Johnson. 2002. Measuring preferences for health care interventions using conjoint analysis. *Health Services Research*, pages 1681–1705.

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

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Vilar, D., J. Xiu, L.F. D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of LREC*.