

# Personalized Machine Translation: Predicting Translational Preferences

Shachar Mirkin\*

IBM Research - Haifa  
Mount Carmel, Haifa  
31905, Israel  
shacharm@il.ibm.com

Jean-Luc Meunier

Xerox Research Centre Europe  
6 chemin de Maupertuis  
38240 Meylan, France  
jean-luc.meunier@xrce.xerox.com

## Abstract

Machine Translation (MT) has advanced in recent years to produce better translations for clients' specific domains, and sophisticated tools allow professional translators to obtain translations according to their prior edits. We suggest that MT should be further personalized to the end-user level – the receiver or the author of the text – as done in other applications. As a step in that direction, we propose a method based on a recommender systems approach where the user's preferred translation is predicted based on preferences of similar users. In our experiments, this method outperforms a set of non-personalized methods, suggesting that user preference information can be employed to provide better-suited translations for each user.

## 1 Introduction

Technologies are increasingly personalized, accommodating their behavior for each user. Such personalization is done through user modeling where the goal is to “get to know” the user. To that end, personalization is based on users' attributes, such as demographics (gender, age etc.), personalities, and preferences. For example, in Information Retrieval, results are customized according to the user's information and search history (Speretta and Gauch, 2005), performance of Automatic Speech Recognition substantially improves when adapted to a specific speaker (Neumeyer et al., 1995), and Targeted Advertising makes use of the user's location and prior purchases (Kölmel and Alexakis, 2002).

Personalization in machine translation has a somewhat different nature. Providers of MT tools and services offer means to “customize” or “personalize” the translation engine for each client, mostly through domain adaptation techniques, and a great deal of effort is made to make the human-involved translation process more efficient (see Section 2.2). Most of the focus, though, goes to customization for companies or professional translators. We argue that *Personalized Machine Translation* (PMT below) should and can take the next step and directly address individual end-users.

\*This work was done while the first author was at Xerox Research Centre Europe.

The difficulty to objectively determine whether one (automatic) translation is better than another has been repeatedly revealed in the MT literature. Our conjecture is that one reason is individual preferences, to which we refer as *Translational Preferences* (TP). TP come into play both when the alternative translations are all correct, and when each of them is wrong in a different way. In the former case, a preference may be a stylistic choice, and in the latter, a matter of comprehension or a selection of the least intolerable error in one's opinion. For instance, one user may prefer shorter sentences than others; she may favor a more formal style, while another would rather have it casual. A user could be fine with some reordering errors but be more picky concerning punctuations. One user will not be bothered if some words are left untranslated (perhaps because the source language belongs to the same language family as the target language that he speaks), while another will find it utterly displeasing. Such differences may be the result of the type of translation system being employed (e.g. syntax- vs. phrased-based), the specific training data or many other factors. On the user's side, a preference may be attributed, for example, to her mother tongue, her age or her personality.

Two aspects of end-user PMT may be considered: (i) Personalized translation of texts written by a specific user, and (ii) PMT to provide better translations for a specific reader. In this work we address the second task, aiming to identify translations each user is more likely to prefer.<sup>1</sup> Specifically, we consider a setting where at least two MT systems are available, and the goal is to predict which of the translation systems the user would choose, assuming we have no knowledge about her preference between them. Benchmarking the systems in advance with respect to a reference set, or estimating the quality of the translations (Specia et al., 2009) are viable alternatives for translation selection; these, however, are not personalized to the target user. Instead, we employ a user-user Collaborative Filtering approach, common in Recommender Systems, which we map to the TP prediction task.

We assess this approach using a collection of user rankings of MT systems from a shared translation task

<sup>1</sup>In (Mirkin et al., 2015) we investigate the first task, assessing whether the author's demographic and personality traits are preserved over machine translation.

(see Section 3). Our results show that the personalized method modestly, but consistently, outperforms several other approaches that rank the systems in general, disregarding the specific user. We consider this as an indication that user feedback can be employed towards a more personalized approach to machine translation.

## 2 Background

### 2.1 Collaborative filtering

Collaborative filtering (CF) is a common approach employed by recommender systems for suggesting to users items, such as books or movies. A recommender system may simply suggest to all users the most popular items; often, however, the recommendations are personalized for each individual user to fit her taste or preferences. User-user CF relies on community preferences. The idea is to recommend to the user items that are liked by users similar to her, as manifested, for example, by high rating. Similar users are those that agree with the current user on previously-rated items. In  $k$ -nearest-neighbors CF, a user is typically represented by a vector of her preferences, where each entry of the vector is, e.g., a rating of a movie.  $k$  similar users are then identified by measuring the similarity between the users' vectors. Cosine similarity is a popular function for that purpose, and we also use it in our work (Resnick et al., 1994; Sarwar et al., 2001; Ricci et al., 2011). An alternative to cosine, Pearson's correlation coefficient (Pearson, 1895), allows addressing different rating patterns across users. In comparison to cosine, here vector entries are normalized with respect to the user's average rating. In our case, such normalization is not very meaningful since the entries of the users vectors represent comparisons rather than absolute ratings, as will be made clear in Section 4. Nevertheless, we have experimented with Pearson correlation as well, and found no advantage in using it instead of cosine.

### 2.2 Customization, personalization and adaptation in MT

Various means of customization and personalization are available, in both academic and commercial MT. Many of them target the company, rather than the individual user, and much of the effort is invested in designing tools for professional translators, aiming to improve their productivity, through intelligent Computer Aided Translation (CAT).

Domain adaptation methods are commonly used to adapt to the topic, the genre and even the style of the translated material. Using the company's own corpora is one of the simplest techniques to do so, but many more approaches have been proposed, including data-selection (Axelrod et al., 2011; Gascó et al., 2012; Mirkin and Besacier, 2014), mixture models (Foster and Kuhn, 2007) and table fill-up (Bisazza et al., 2011). Clients can utilize their own glossaries (Federico et al., 2014), corpora (parallel or monolingual) and translation memories (TM), either shared or private ones

(Caskey and Maskey, 2014; Federico et al., 2014). Through Adaptive and Interactive MT (Nepveu et al., 2004), the system learns from the translator's edits, in order to avoid repeating errors that have already been corrected. Post-editions can continuously be added to the translator's TM or be used as additional training material, for tighter adaptation to the domain of interest, through batch or incremental training.

### 2.3 User preferences in MT

Many tasks that require annotation by humans are affected by the annotator and not only by the item being judged. Metrics for inter-rater reliability or inter-annotator agreement, such as Cohen's Kappa (Cohen, 1960), help measuring the extent to which annotators disagree. Disagreement may be due to untrained or inattentive annotators, a result of a task that is not well defined, or when there is no obvious "truth". Such is the case with the evaluation of translation quality – it is not always straightforward to tell whether one translation is better than another. A single sentence can be translated in multiple correct ways. The decision becomes even harder when the translations are automatically produced and are imperfect: Is one error worse than another? The answer is in the eye of the beholder. MT papers regularly report rather low Kappa levels, even when measured on simpler tasks, such as short segments (Macháček and Bojar, 2015).

Turchi et al. (2013) refer to the issue of "subjectivity" of human annotators. They address the task of binary classification of "good" vs. "bad" translations, and show that relying on human annotation for training a binary quality estimator is less effective than using automatically-generated labels. This subjectivity is exactly what we are after. We treat it as a *preference*, trying to identify the systems or specific translations which the user subjectively prefers.

Kichhoff et al. (2012) analyze user preferences with respect to MT errors. They show that some types, e.g. word order errors, are the most dis-preferred by users, and that this is a more important factor than the number of errors. While very relevant for our research, their analysis is aggregated over all users participating in the study, and is not focusing on individuals' preferences.

## 3 Data

In this work we used the data provided for the MT Shared Task in the 2013 Workshop on Statistical Machine Translation (WMT) (Bojar et al., 2013).<sup>2</sup> This data was of a particularly large scale, with crowd-sourced human judges, either volunteer researchers or paid Amazon Turkers. For each source sentence, a judge was presented with the source sentence itself, a reference translation, and the outputs of five machine translation systems. The five systems were randomly selected from the pool of participating systems,

<sup>2</sup><http://www.statmt.org/wmt13/translation-task.html>

and were anonymized and randomly-ordered when presented to the judge. The judge had to rank the translations, with ties allowed (i.e. two system can receive the same ranking). Hence, each annotation point provided with 10 pairwise rankings between systems. Translations of 10 language pairs were assessed, with 11 to 19 systems for each pair. In total, over 900K non-tied pairwise rankings were collected. The Turkers’ annotation included a control task for quality assurance, rejecting Turkers failing more than 50% of the control points. The inter-annotator score showed on average a fair to moderate level of agreement.

## 4 Translational preferences with collaborative filtering

Our method, denoted CTP (*Collaborative Translational Preferences*), is based on a  $k$ -nearest-neighbors approach for user-user CF. That is, we predict the translational preferences of a user based on those of similar users. In our setting, a user preference is the choice between two translation systems – which system’s translations does the user prefer. Given two systems (or models of the same system) we wish to predict which one the user would prefer, without assuming the user has ever expressed her preference between these two specific systems. It is important to emphasize that the method presented here considers the users’ overall preferences of systems, and does not regard the specific sentence that is being translated. In future work we intend to make use of this information as well.

### 4.1 Representation

As mentioned in Section 3, each annotation consists of a ranking of five systems. From that, we extract pairwise rankings for every pair of systems that were ranked for a given language pair. For each user  $u \in \mathcal{U}$  (where  $\mathcal{U}$  are all users who annotated the language pair), we create a user-preference vector,  $p_u$ , that contains an entry for each pair of translation systems. Denoting the set of systems with  $\mathcal{S}$ , we have  $\frac{|\mathcal{S}| \cdot (|\mathcal{S}|-1)}{2}$  system pairs. E.g., for Czech-English, with 11 participating systems, the user vector size is 55. Each entry  $(i, j)$  of the vector is assigned the following value:

$$p_u^{(i,j)} = \frac{w_u^{(i,j)} - l_u^{(i,j)}}{w_u^{(i,j)} + l_u^{(i,j)}} \quad (1)$$

where  $w_u^{(i,j)}$  and  $l_u^{(i,j)}$  are the number of wins and losses of system  $s_i$  vs. system  $s_j$  as judged by user  $u$ .<sup>3</sup>

With this representation, a user vector contains values between  $-1$  (if  $s_i$  always lost to  $s_j$ ) and  $1$  (if  $s_i$  always won). If the user always ranked the two systems identically, the value is  $0$ , and if she has never evaluated the pair, the entry is regarded as a missing value ( $NA$ ). Altogether, we have a matrix of users by system pairs, as depicted in Figure 1.

<sup>3</sup>We have also considered including ties in the denominator of the equation; discarding them was found superior.

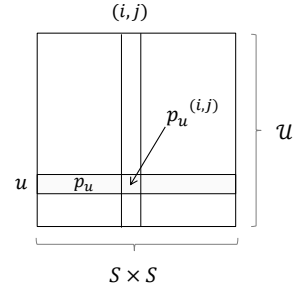


Figure 1: The user-preferences matrix.

### 4.2 Finding similar users

Given a user preference to predict for a pair of systems  $(s_i, s_j)$ , we compute the similarity between  $p_u$  and each one of  $p_{u'}$  for all other  $u' \in \mathcal{U}$ . In our experiments we used cosine as the similarity measure. The  $k$  most-similar-users ( $MSU$ ) are then selected. To be included in  $MSU(u)$ , we require that  $u$  and  $u'$  have judged at least 2 common system pairs.

### 4.3 Preference prediction

Given the similarity scores, to predict the user’s preference for the target system pair, we compute a weighted average of the predictions of the users in  $MSU(u)$ .

We include in the average only users with similarity scores above a certain positive threshold (0.05). We then require that a minimum number of users meet the above criteria of common annotations and minimum similarity (we used 5). If not enough such similar users are found, we turn to a fallback, where we use the non-weighted average preference across all users (AVPF presented in Section 5).<sup>4</sup> The prediction is then the sign of the weighted average. A positive value means  $s_i$  is the preferred system; a negative one means it is  $s_j$ , and a zero is a draw. In our evaluation we compare this prediction to the sign of the actual preference of the user,  $p_u^{(i,j)}$ . Formally, CTP computes the following prediction function  $f$  for a given user  $u$  and a system pair  $(s_i, s_j)$ :

$$f_{CTP}(u)^{(i,j)} = \text{sign}\left(\frac{\sum_{u'} p_{u'}^{(i,j)} \cdot \text{sim}(u, u')}{\sum_{u'} \text{sim}(u, u')}\right) \quad (2)$$

where  $u' \in MSU(u)$  are the most similar users (the nearest neighbors) of  $u$ ;  $p_{u'}^{(i,j)}$  are the preferences of user  $u'$  for  $(s_i, s_j)$  and  $\text{sim}(u, u')$  is the similarity score between the two users.<sup>5</sup>

## 5 Experiments and results

### 5.1 Evaluation methodology

In our experiments we try to predict which one of two translation systems would be preferred by a given user.

<sup>4</sup>The fallback was used 0.1% of the times.

<sup>5</sup>The denominator is not required as long as we predict only the sign since all used similarity scores are positive. We keep it in order to obtain a normalized score that can be used for other decisions, e.g. ranking multiple systems.

We evaluate our method, as well as several other prediction functions, when compared with the user’s pairwise system preference according to the annotation –  $p_u^{(i,j)}$ , shown in Equation 1. For each user this is an aggregated figure over all her pairwise rankings for the pair, determining the preferred system as the one chosen by the user (i.e. ranked higher) more times.

We conduct a leave-one-out experiment. For each language pair, we iterate over all non-*NA* entries in the user-preferences matrix, remove the entry and try to predict it. User similarity scores are re-computed for each evaluation point, to ensure they do not consider the target pair. The “gold” preference is positive when the user prefers  $s_i$ , negative when she prefers  $s_j$  and 0 when she has no preference between them. Hence, each of the assessed methods is measured by the accuracy of predicting the sign of the preference.

## 5.2 Non-personalized methods

We compare CTP to the following prediction methods:

**Always  $i$  (ALI)** This is a naïve baseline showing the score when always predicting that system  $i$  wins. Note that the baseline is not simply 50% due to ties.

**Average rank (RANK)** Here, two systems are compared by the average of their rankings across all annotations ( $r \in \{1, 2, 3, 4, 5\}$ ):

$$f_{\text{RANK}}(u)^{(i,j)} = \text{sign}(\bar{r}_j - \bar{r}_i) \quad (3)$$

$\bar{r}_j$  and  $\bar{r}_i$  are the average ranks of  $s_j$  and  $s_i$  respectively. Since a smaller value of  $r$  corresponds to a higher rank, we subtract the rank of  $s_i$  from  $s_j$  and not the other way around. This way, if for instance,  $s_i$  is ranked on averaged higher than  $s_j$ , the prediction would be positive, as desired.

**Expected (EXPT)** This metric, proposed by Koehn (2012) and used by Bojar et al. (2013) in order to rank the participating systems in the WMT benchmark, compares the *expected wins* of the two systems. Its intuition is explained as follows: “If the system is compared against a randomly picked opposing system, on a randomly picked sentence, by a randomly picked judge, what is the probability that its translation is ranked higher?” The expected wins of  $s_i$ ,  $e(s_i)$ , is the probability of  $s_i$  to win when compared to another system, estimated as the total number of wins of  $s_i$  relative to the total number of comparisons involving it, excluding ties, and normalized by the total number of systems excluding  $s_i$ ,  $|\{s_k\}|$ :

$$e(i) = \frac{1}{|\{s_k\}|} \sum_{k \neq i} \frac{w^{(i,k)}}{w^{(i,k)} + l^{(i,k)}} \quad (4)$$

where  $w^{(i,k)}$  and  $l^{(i,k)}$  are summed over all users. The preference prediction is therefore:

$$f_{\text{EXPT}}(u)^{(i,j)} = \text{sign}(e(i) - e(j)) \quad (5)$$

RANK and EXPT predict preferences based on a system’s performance in general, when compared to all other systems. We propose an additional prediction function for comparison which uses only the information concerning the system pair under consideration.

**Average user preference (AVPF)** This method takes into account only the specific system pair and averages the user preferences for the pair. Formally:

$$f_{\text{AVPF}}(u)^{(i,j)} = \text{sign}\left(\frac{\sum_{u'} p_{u'}^{(i,j)}}{|\{u'\}|}\right) \quad (6)$$

where  $u' \neq u$ , and  $\{u'\}$  are all users except  $u$ .

This method can be viewed as a non-personalized version of CTP, with two differences:

- (1) It considers all users, and not only similar ones.
- (2) It does not weight the preferences of the other users by their similarity to the target user.

## 5.3 Results

Table 1 shows the results of an experiment comparing the performance of the various methods in terms of prediction accuracy. Figure 2 shows the micro-average scores, when giving each of the 97,412 test points an equal weight in the average. CTP outperforms all others for 9 out of 10 language pairs, and in the overall micro-averaged results. The difference between CTP and each of the other metrics was found statistically significance with  $p < 5 \cdot 10^{-6}$  at worse, as measured with a paired Wilcoxon signed rank test (Wilcoxon, 1945) on the predictions of the two methods. The significance test captures in this case the fact that the methods disagreed in many more cases than is visible by the score difference.

Our method was found superior to all others also when computing macro-average, taking the average of the scores of each language pair, as well as when the ties are included in the computation of  $p_u$ .

The parameters with which the above results were obtained are found within the method’s description in Section 4. Yet, in our experiments, CTP turned out to be rather insensitive to their values. In this experiment we used a global set of parameters and did not tune them for each language pair separately. It is reasonable to assume that such tuning would improve results. For instance, choosing  $k$ , the number of users to include in the average, depends on the total number of users. E.g., for *en-es*, where there are only 57 users in total, reducing  $k$ ’s value from 50 to 25, improves results of CTP from 62.6% to 63.2%, higher than all other methods (whose scores are not affected).

Specifically in comparison to AVPF, weighting by the similarity scores was found to be a more significant factor than selecting a small subset of the users. This may not come as a surprise, since less similar users that are added to  $MSU(u)$  have a smaller impact on the final decision since their weight in the average is smaller.

Lang.	<i>f</i>	Acc.	Lang.	<i>f</i>	Acc.	Lang.	<i>f</i>	Acc.	Lang.	<i>f</i>	Acc.	Lang.	<i>f</i>	Acc.
cs-en	ALI	31.6	de-en	ALI	41.7	es-en	ALI	35.5	fr-en	ALI	35.0	ru-en	ALI	43.5
	RANK	62.9		RANK	62.6		RANK	61.0		RANK	61.3		RANK	57.6
	EXPT	63.5		EXPT	62.6		EXPT	61.2		EXPT	61.2		EXPT	57.8
	AVPF	63.5		AVPF	62.6		AVPF	61.4		AVPF	61.2		AVPF	56.6
	CTP	<b>64.4</b>		CTP	<b>63.5</b>		CTP	<b>63.0</b>		CTP	<b>61.8</b>		CTP	<b>58.2</b>
en-cs	ALI	36.2	en-de	ALI	42.0	en-es	ALI	35.9	en-fr	ALI	35.0	en-ru	ALI	44.6
	RANK	67.8		RANK	67.2		RANK	62.3		RANK	65.0		RANK	70.2
	EXPT	67.9		EXPT	66.9		EXPT	<b>63.0</b>		EXPT	65.1		EXPT	72.1
	AVPF	67.4		AVPF	66.5		AVPF	61.5		AVPF	64.4		AVPF	71.4
	CTP	<b>68.2</b>		CTP	<b>67.6</b>		CTP	62.6		CTP	<b>65.3</b>		CTP	<b>72.4</b>

Table 1: Results in accuracy percentage for the 10 language pairs, including the languages: Czech (*cs*), English (*en*), German (*de*), Spanish (*es*), French (*fr*) and Russian (*ru*). The best results is in bold. The difference between CTP and each of the other methods is highly statistically significant. Figure 2 shows a micro-average of these results.

One weakness of CTP, as well as of other methods, is that it poorly predict ties. In the above experiment, approximately 13.5% of the preferences were 0, none of them was correctly identified. Our analysis showed that numerical accuracy is not the main cause; setting any prediction that is smaller than some values of  $|\epsilon|$  to 0 was not found helpful. Arguably, ties need not be predicted, since if the user has no preference between two systems, any choice is just as good. Still, we believe that better ties prediction could lead to general improvement of our method and we wish to address it in future work.

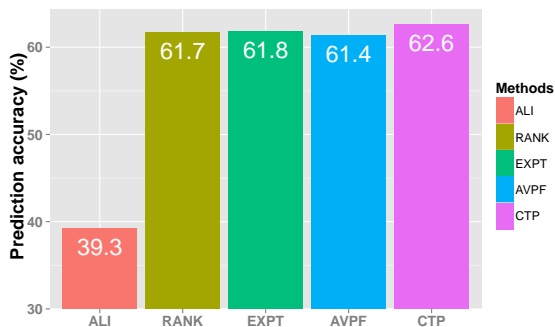


Figure 2: Micro-average over all 97,412 test points.

## 6 Discussion

We addressed the task of predicting user preference with respect to MT output via a collaborative filtering approach whose prediction is based on preferences of similar users. This method predicts TP better than a set of non-personalized methods. The gain is modest in absolute numbers, but the results are highly statistically significant and stable over parameter values.

We consider this work as a step towards more personalized MT. This line of research can be extended in multiple ways. First and foremost, as mentioned, we did not consider the actual content of the sentences, but rather identified a general preference for one system over another. It is plausible, however, that one system is better – from the user’s perspective – at translating one type of text, while another is preferred for other texts. Taking the actual texts into account seems therefore es-

sential. Content-based methods for recommender systems may be useful for this purpose. Another factor that may be affecting preferences is translation quality: when compared translations are all poor, preferences play a less significant role. Hence, it may be informative to assess TP prediction separately across different levels of translation quality.

Large parallel corpora are typically required for training reasonable statistical translation models. Yet, parallel corpora, and even more so in-domain ones, are hard to gather. It is virtually impossible to find a user-specific parallel corpus, and methods for monolingual domain adaptation are easier to envisage if one wishes to address author-aware PMT (the first PMT task mentioned in Section 1). Collecting user feedback is another challenge, especially since most end-users do not speak the source language. For that and other reasons, it currently seems more feasible to collect preference information from professional translators, explicitly or implicitly. Yet, in this research we aim at end-users rather than translators whose preferences are often driven by the ease of correction more than anything else. We believe that one way to tackle this issue is to exploit other kinds of feedback, from which we can infer user preferences and similarity. Online MT providers are recently collecting end-user feedback for their proposed translations which may be useful for TP prediction. For instance, in early 2015 Facebook introduced a feature letting users rate (Bing) translations, and Google Translate asks for suggested improvements. We are hopeful that such data becomes publicly available. Nevertheless, it remains unlikely to obtain feedback from each and every user. A potential direction for both corpora and feedback collection is personalizing models and identifying preferences for groups of users based on socio-demographic traits, such as gender, age or mother tongue, or based on (e.g. Big 5) personality traits. These can even be inferred by automatically analyzing user texts.

## Acknowledgments

We wish to thank Hervé Déjean and the EMNLP reviewers for their valuable feedback on this work.

## References

- Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, California, USA.
- Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Sasha P. Caskey and Sameer Maskey. 2014. Translation cache prediction, August 12. US Patent 8,805,672.
- Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.
- Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. The matecat tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.
- George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, Prague, Czech Republic, June.
- Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–161, Avignon, France, April. Association for Computational Linguistics.
- Katrin Kirchhoff, Daniel Capurro, and Anne Turner. 2012. Evaluating user preferences in machine translation using conjoint analysis. *Proceedings of the European Association of Machine Translation*, 16:119–126.
- Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 179–184, Hong Kong.
- Bernhard Kölmel and Spiros Alexakis. 2002. Location based advertising. In *Proceedings of the First International Conference on Mobile Business, Athens, Greece*.
- Matouš Macháček and Ondřej Bojar. 2015. Evaluating machine translation quality using short segments annotations. *The Prague Bulletin of Mathematical Linguistics*, No. 103:85–110, April.
- Shachar Mirkin and Laurant Besacier. 2014. Data selection for compact adapted SMT models. In *Proceedings of the eleventh biennial conference of the Association for Machine Translation in the Americas (AMTA-2014)*, Vancouver, Canada, Oct.
- Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal. Association for Computational Linguistics.
- Laurent Nepveu, P Langlais, G Lapalme, and George Foster. 2004. Adaptive language and translation models for interactive machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–197, Barcelona, Spain.
- Leonardo Neumeyer, Ananth Sankar, and Vassilios Digalakis. 1995. A comparative study of speaker adaptation techniques. In *Fourth European Conference on Speech Communication and Technology, EUROSPEECH 1995, Madrid, Spain, September 18-21, 1995*.
- Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242.
- Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM.
- Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook.
- Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM.
- Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

- Mirco Speretta and Susan Gauch. 2005. Personalized search based on user search histories. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 622–628. IEEE.
- Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgments in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 240–251, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, December.