Motivating Personality-aware Machine Translation

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

Scott Nowson, Caroline Brun, Julien Perez
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
firstname.surname@xrce.xerox.com

Abstract

Language use is known to be influenced by personality traits as well as by socio-demographic characteristics such as age or mother tongue. As a result, it is possible to automatically identify these traits of the author from her texts. It has recently been shown that knowledge of such dimensions can improve performance in NLP tasks such as topic and sentiment modeling. We posit that machine translation is another application that should be personalized. In order to motivate this, we explore whether translation preserves demographic and psychometric traits. We show that, largely, both translation of the source training data into the target language, and the target test data into the source language has a detrimental effect on the accuracy of predicting author traits. We argue that this supports the need for personal and personality-aware machine translation models.

1 Introduction

Computational personality recognition is garnering increasing interest with a number of recent workshops exploring the topic (Celli et al., 2014; Tkalčič et al., 2014). The addition of personality as target traits in the PAN Author Profiling challenge in 2015 (Rangel et al., 2015) is further evidence. Such user modeling – when performed on text – is built on a long-standing understanding that language use is influenced by socio-demographic characteristics such as age, gender, education level or mother tongue and personality traits like agreeableness or openness (Tannen, 1990; Pennebaker et al., 2003).

In this work we explore *multilingual* user modelling. The motivation is not only to enable modeling in multiple languages, but also to enable modeling multilingual users who may express different sides of their personality in each language. One way to address multilinguality in this context is to create models separately in each language, and then fuse the resulting models. However, labelled data of this nature, particularly in non-English languages, is often not available. Personality labelling is time consuming, requiring the completion of psychometric questionnaires which may be considered invasive by many. An alternative is the use of machine translation (MT) to bootstrap corpora in resource poor languages, and to translate the user's content into a single language before modeling. Translated text, either manually or automatically generated, is known to have different characteristics than native text. Yet, MT was shown to be of use within traditional NLP tasks such as sentiment analysis (Balahur and Turchi, 2012). We explore the utility of MT for classification of demographic and personality traits.

MT models, even domain-specific, are user-generic. Thus, the linguistic signals of user traits which are conveyed in the original language may not be preserved over translation. In other words, the attributes on which we wish to rely for modelling may be lost. This concern is perhaps most observable with gender, a trait of the speaker that is encoded in the morphology of many languages, though not in English. Gendered translation was the topic of research for many years. However, the gender of the author is largely ignored by MT systems, and specifically statistical ones, that would often arbitrarily (or rather statistically-based) translate into one gender form or another. Other demographic and personality traits have not yet been investigated.

One way to address this concern is *personalized* translation, or *author-aware* translation.[1] The first step toward this goal would be to consider the author traits in the model. Such an approach has already shown to be useful for several NLP tasks (Volkova et al., 2013; Hovy, 2015). However, before embarking on this challenging task, we explore if the above concerns are founded by addressing the research question: does MT has an impact on the classification of demographic and personality traits?

2 Background

Oberlander and Nowson (2006) motivated their study of computational personality recognition by arguing that automatically understanding an author's personality would permit the personalization of sentiment analysis. Such personalized NLP has recently been

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

[1] In this work we investigate MT awareness of the author; in (Mirkin and Meunier, 2015) we address the task of reader-aware MT.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1102–1108, Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

explored by Volkova et al. (2013). They incorporated age and gender features for sentiment analysis, and show improvements in three different languages. Hovy (2015) extends this work to other languages and NLP tasks. Using demographically-informed word embeddings, they show improvements in sentiment analysis, topic classification and trait detection. None of these works, however, addressed cross-lingual issues.

Yet, personality projection goes beyond automatic detection of traits – there is also human perception to be considered. The casual reader may not be aware of personality related linguistic cues. Yet, studies have shown that traits can be reliably detected following cold readings of texts from unknown authors (Mehl et al., 2006; Gill et al., 2012) without such explicit knowledge. Although personality projection in different languages is under-explored, it has been shown that the relationship between language use and personality traits varies between domains (Nowson and Gill, 2014). Thus, while it would seem that there are cues which translate directly between languages, this may not always be the case. In English, for example, women tend to use first-person pronouns such as "*I*" more than men (Newman et al., 2008); but this does not guarantee a gender-based usage difference for, say, "*je*" in French. Furthermore, what happens with more subtle, language-specific indicators of personality? For instance, Nowson (2006) showed that use of contractions (e.g. *don't* vs. *do not*) is a marker for the Agreeableness trait. These different forms do not naturally translate into other languages. It is doubtful that even a human translator would always pay attention to such subtleties. In investigating whether these cues are preserved when a text is translated, we are also beginning to address the question of consistency in cues between languages.

MT systems do not explicitly consider demographic or personality traits. Instead, they often exploit "in-domain" data to create translation models that are adapted for the domain of interest (Lu et al., 2007; Foster et al., 2010; Axelrod et al., 2011; Gong et al., 2012; Mirkin and Besacier, 2014). The term "domain" has a wide interpretation in the MT literature and may refer to topic, dialect, genre or style (Chen et al., 2013). However, to the best of our knowledge, MT domain adaptation does not extend to consider demographic or personality traits of the author. Gender in translation has been researched extensively; in human translation studies, it has been shown that the gender of translators impact the translation. In SMT, phrase-based models (Koehn et al., 2003) can correctly pick-up translations of gender-inflected words, and rule-based MT systems and factored models (Koehn and Hoang, 2007) provide more explicit ways for gender translation. Yet, most SMT systems are unaware of the gender of the author, neither in the training nor in the test data, and are therefore unable to adapt their translation beyond the local inflectional level; in particular when no such evidence exists, as in English. To a much greater extent, this is the case with other demographics, such as age, and with personality traits.

3 Methodology

3.1 Hypothesis

The hypothesis of our broader vision is that personalized MT or author-aware translation is an important necessity. We believe the human understanding of translated text (of its explicit and implicit meanings, of its author and of the full context) would be improved if author traits are better conveyed.

In order to motivate this future work, this paper explores a supporting hypothesis: that author traits are not conveyed accurately under machine translation. We assess this by investigating whether trait detection performs as well on translated data as on native text.

3.2 Experimental Framework

To explore our hypothesis, we require data in multiple languages which is labelled with socio-demographic or personality traits. Using English as the base language (as typically the most resource-rich language in NLP studies), we perform three comparative experiments on several non-English ("foreign") corpora. In these experiments we train a classification model:

1. Using only foreign language data. This provides a baseline, as no translated data is used.
2. Augmenting the foreign training data with English data translated into the foreign language. Here, the goal is to assess a scenario where translations from a resource-rich language supplement scarce training data in the foreign language, under the assumption that more training data can be beneficial.
3. Translating the foreign test data into English and classifying it using a model trained on the English data. This allows us to explore another practical scenario, where an English model already exists and we wish to use it to classify data from another language for which we do not have a robust model.

For this task we use the data from the 2015 PAN workshop (Rangel et al., 2015) which is labelled for author gender and personality traits. For more details see Section 4.1. We also wish to explore if any affect was due strictly to the use of MT or to translation (or language change) in general. The PAN corpus is multi-lingual but does not contain parallel data. Such parallel corpora, however, are not typically labelled with the type of author information we wish to investigate. Therefore we required such a corpus to which we could easily add labels. For these we used a selection of TED talks which we labelled for gender (see Section 4.2). The full details of our approach to text processing, translation and classification can be found in our technical paper at the PAN workshop (Nowson et al., 2015); in the interests of space a compressed version is presented here.

3.3 Preprocessing and feature extraction

We use the multilingual parser described by Ait-Mokhtar et al. (2001) to preprocess the texts and extract a wide range of features. The parser has been customized to handle social media data, e.g. by detecting hashtags, mentions, and emoticons. For English, we have integrated a normalization dictionary by Han et al. (2012) in the preprocessing. The English and French grammars also include a polarity lexicon to recognize sentiment bearing words or expressions. The features we extract include: 1-, 2-, 3-grams of surface, normalized and lemmatized forms; part-of-speech tagged forms, and n-grams of POS; named entities (places, persons, organization, dates, time expressions), emoticons, hashtags, mentions and URLs.

3.4 Learning framework

To train classification models we first prune features with a frequency threshold. Next, the remaining set of features is compressed using truncated singular value decomposition (SVD). SVD (Golub and Reinsch, 1970) is a widely used technique in sparse dataset situations. This method copes with noise present in the data by extracting the principal dimensions describing the data and projecting the data to a latent space. In the truncated version, a low-rank approximation, all but the top-k dimensions are removed. The result is a dense, low-dimension representation of the data. Finally, ensemble models (Schapire, 1990; Dietterich, 2000) are used to predict trait values: for gender, we use the majority vote of 10 classifiers; for each personality trait, we use the mean of 10 regression estimators.

3.5 Machine translation models

We created standard machine translation models between English and each one of Spanish, Italian, French and Dutch. The details are described below.

Parallel corpora We wished to use the same setting for all language pairs. To that end, we chose parallel corpora that are available for all of them, namely Europarl (Koehn, 2005)[2] and WIT3 (Cettolo et al., 2012), from the IWSLT 2014 evaluation campaign (Cettolo et al., 2014). WIT3, consisting of spoken-language transcripts, represents an in-domain corpus for the TED dataset and a "near-domain" for PAN. The data consisted of approximately 2 million parallel sentences for each language pair, with 50 million tokens for each language. The Europarl corpus comprised more than 90% of that data. The two corpora were concatenated to create the training data for the MT models.

Translation System Moses (Koehn et al., 2007), an open-source phrase-based MT system,[3] was used to train translation models and translate the data.

[2]Version 7, `www.statmt.org/europarl`

[3]We used version 3.0, downloaded on 16 Feb 2015 from `www.statmt.org/moses`.

Preprocessing We used the standard Moses tools to preprocess the data, including tokenization, lowercasing and removal of sentence pairs where at least one of the sentences is empty or longer than 80 tokens.

Recasing and Language models We used SRILM (Stolcke, 2002) version 1.7.1 to train 5-gram language models on the target side of the parallel corpus, with modified Kneser-Ney discounting (Chen and Goodman, 1996). A recasing model was trained from the same corpus, with a 3-gram KenLM language model (Heafield, 2011).

Tuning We tuned the translation models using MERT (Och, 2003), using the development set of the above mentioned campaign (*dev2010*), consisting of 887 sentence pairs for each language pair.

Translation and post-processing Each of the tweets of the PAN training set was preprocessed in the same fashion as the training data. It was then translated with the trained model of the corresponding language pair, and finally underwent quick post-processing, namely recasing and detokenization.

4 Data

Personality-tagged datasets in multiple languages are scarce. We used two datasets, with content from twitter and TED talks, as described in this section.

4.1 PAN

The first corpus we used was the data of the PAN 2015 Author Profiling task (Rangel et al., 2015), drawn from Twitter (PAN15). For each user, the data consists of tweets (average $n = 100$) and gold standard labels: gender (Male or Female), and personality. The labels are provided by the author, with scores on five traits being calculated via self-assessment responses to the short Big 5 test, BFI-10 (Rammstedt and John, 2007)), then normalized between -0.5 and +0.5. Table 1 shows the volume of data per language for the training set.

Language	Authors	Tweets
English (en)	152	14166
Spanish (es)	100	9879
Italian (it)	38	3687
Dutch (nl)	34	3350

Table 1: Number of authors and tweets across the four languages of the PAN dataset.

4.2 TED

The PAN15 data allows us to assess personality projection in multilingual data. In addition to exploring automatic translation, we wish to compare with manual translation. We turned to TED talks[4] for such comparative evaluation. We chose the English-French lan-

[4]`www.ted.com`

guage pair, because French is not a language in PAN15, but also due to the difficulties in obtaining such data, as described below.

4.2.1 TED English-French

We use data of the MT track of the IWSLT 2014 Evaluation Campaign, which includes parallel corpora from transcripts of TED talks. The English-French (*en-fr*) corpus consists of 1415 talks, with approximately 190k sentence pairs and 3 million tokens for each of the source and target sides (before preprocessing). We annotated the gender of each speaker with a simple web interface. Any talk with multiple speakers or where the majority is not a speech (e.g. a performance) is discarded. After discarding 59 talks, 1012 (75%) were annotated as male and 344 (25%) as female.[5]

4.2.2 TED French-English

The WIT3 data seems to also include data in the *fr-en* direction. However, in practice, TED hosts only talks in English and all foreign to English corpora were collected from the translated versions of the site. We therefore turned to TEDx[6] for *fr-en* data. TEDx are independent TED-style events, often including talks in languages other than English. Unlike *en-fr*, there is no easily accessible parallel data available for *fr-en*, where the source is native French. We applied the following procedure to collect the necessary data. We used the Google YouTube Analytics API[7] to search for videos of talks in French. We have extracted the list of TEDx events in France and their dates via `www.tedxenfrance.fr`.[8] Each event-name and year is used as a query in YouTube, e.g. "TEDx Paris 2011". For each talk, we download the **manual** French and English subtitles, i.e. the transcript and the translation, respectively. These files were annotated using the same process and criteria described above. This resulted with a small corpus (TED61$_{fr-en}$) of 61 talks of which 32 are annotated as male and 29 as female.

TED61$_{en}$ In order to account for any potential effect of length, we created a subset of the *en-fr* corpus, that is of the same size of the *fr-en* dataset. We matched files from the French side of the *en-fr* corpus to each of those in the *fr-en* for gender and length (in tokens). The French *en-fr* files were truncated after the nearest line break to the desired size; the corresponding English *en-fr* files were truncated at the same point.

5 Experiments

It has been shown that standard approaches to gender classification on English texts can be sub-optimal for non-English language data (Ciot et al., 2013). However, state-of-the-art classification results are not our focus; rather, our intention is to understand the impact of translation on classification of socio-demographic and personality traits. Therefore, we fix our models with a set of parameters selected via cross-validation (CV) on the native language: the occurrence threshold is set to 5 and the SVD dimensionality to 500.

5.1 PAN

For each of the three non-English languages of PAN15 we train classification models as explained in Section 3: using the original training data, adding training data translated from English, and translating the test data into English to use the English-trained model.

Results can be seen in Table 2. For the majority of the traits, the native results outperform both translation settings, in some cases by considerable margin. The assumption posited earlier that more training data is be beneficial appears not to have held up in this context. The alternative scenario seems to be doing even worse.

The most distinct results are perhaps the accuracy of gender prediction (for which each corpora is balanced, thus a baseline of 50%). One explanation may be that the translation is done from and into English, which does not express gender via morphology, in contrast to Italian and Spanish. An interesting exception is that adding translated English texts into Dutch considerably improves performance. This may be explained by the lesser expression of gender in Dutch morphology, much like English. In this instance it appears that adding more data – when translation is between two gender-agnostic languages – does indeed help. For both Italian and Dutch, English adds a very substantial amount of data; the outcomes, however, are opposite.

5.2 TED

Though the TED data is currently only labelled for gender, it allows us to make comparisons between manual and machine translation. First we explored if there were gender signals in the English corpus which a classifier could uncover. For this, we performed leave-one-out CV on each of the following three versions: native English, manually and machine translated into French. The results, presented in Table 3, show that some gender signal is lost between manual and machine translation. In the manual translation, the translator, who is aware of the speaker's gender is able to reflect that through morphological and lexical cues, that exist in French much more than in English. The MT's ability to project these features properly was more limited. Note that the results between English and French are not directly comparable, since any text classification on different languages may yield different results.

One interesting observation is the low performance relative to the baseline and that of PAN15. Though we do not discuss this in detail here, we suspect this may be an effect of genre muting (Argamon et al., 2003; Herring and Paolillo, 2006).

Next, we explore another setting: The English corpus is modified to exclude the speakers of the

[5] Annotation data is available at `cm.xrce.xerox.com`
[6] `www.ted.com/watch/tedx-talks`
[7] `developers.google.com/youtube`
[8] Accessed on 23/2/15.

Training	Test	Gender (%)	Extraverted	Stable	Agreeable	Conscientious	Open
en	*en*	80.5	0.029	0.050	0.030	0.021	0.021
es	*es*	**82.8**	**0.023**	**0.035**	0.024	0.024	0.025
es + *en*→*es*	*es*	75.1	0.031	0.042	0.024	**0.021**	**0.020**
en	*es*→*en*	62.6	0.032	0.048	**0.021**	0.027	0.030
it	*it*	**80.0**	**0.009**	0.028	0.020	**0.010**	0.019
it + *en*→*it*	*it*	59.1	0.013	0.028	**0.016**	0.014	**0.016**
en	*it*→*en*	61.7	0.031	0.063	0.020	0.024	0.025
nl	*nl*	67.6	**0.008**	**0.014**	**0.014**	**0.007**	**0.010**
nl + *en*→*nl*	*nl*	**74.0**	0.011	0.032	0.020	0.015	0.012
en	*nl*→*en*	53.2	0.028	0.076	0.018	0.017	0.023

Table 2: Cross-validation results on PAN15 for the settings as per Section 5.1. Gender is measured in accuracy; the remaining traits as mean squared error. Bold highlights the best result. English results are included for comparison.

Corpus	English	→French
Native	63.1	
Manual		66.6
MT		62.7

Table 3: Gender CV accuracy (%) on the English TED dataset, when translated manually and automatically.

Corpus	Accuracy (%)
TED61_{en}	58.3
TED61_{fr-en} (Manual)	60.0
TED61_{fr-en} (MT)	52.1

Table 4: Results when classifying gender on native, manually translated and machine translated English texts, from 61 TEDx and TED talks.

TED61_{en} dataset (leaving $n = 1295$ speakers), and this data is used to train a classification model. We then test our three smaller English datasets on this model: TED61_{en}, TED61_{fr-en} manual translated and TED61_{fr-en} machine translated. The results in this case are more comparable since we use the same model for all datasets and since their sizes are similar. The classification results are presented in Table 4. Again, signal is lost in automatic translation in comparison to manual translation. Interestingly, the manual translation scores higher than the native English, as if the translators are adding more gender indications to the text. Further analysis is required to clarify whether this is indeed a consequence of the manual translation or an artifact of the setting.

Author-aware translation may be viewed as a human-centric domain adaptation task: we can consider the two genders as two different domains, and apply domain adaptation techniques to train a better-suited model for each one. To assess this approach, we conducted a set of experiments with standard domain adaptation techniques for *en-fr*, including: separating the translation models and the language models by gender in various configurations, using only the target gender's training data from WIT3 (on top of the Europarl data), and separating tuning sets by gender. We split the IWSLT test sets by gender, and applied on each part the respective gender's model before concatenating the translations to compute a BLEU (Papineni et al., 2002) score. Unfortunately, none of these models showed a significant improvement, if at all, in comparison to our baseline that used both genders together. This suggests that alternative methods should be used for our task. We cannot say, however, that these results are conclusive; specifically, one difficulty in our experiments was obtaining enough female data, due to the relative small number of female speakers in WIT3.

6 Discussion

We are interested in understanding the impact which the consideration of author traits might have on automatic translation, in order to preserve projection of those traits in a target language. However, it is first necessary to understand the inverse: the effect of current translation approaches on the computational recognition of these traits. In the initial studies reported here we have explored two corpora: one of social media data; one of scripted speeches. Although linguistic signals of traits are weaker in the latter case, so far it appears that machine translation is detrimental to the automatic recognition of these traits. Though we have tried to account for as many confounding factors in this work as we could – particularly the availability of data – naturally there are still some open questions, and some obvious next steps. We fixed the learning parameters across languages and traits for comparative reasons, but would independent optimization provide better results? What is the impact of the translation quality on the subsequent classification performance? We would also like to understand the true relationship between linguistic features and traits across languages, along with how native speakers naturally observe these traits. Overall, however, we are encouraged to pursue our goal of personalized machine translation.

Acknowledgments

We would like to gratefully acknowledge the comments and feedback we received from the EMNLP reviewers.

References

Salah Ait-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2001. A multi-input dependency parser. In *Proceedings of IWPT*.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 52–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabio Celli, Bruno Lepri, Joan-Isaac Biel, Daniel Gatica-Perez, Giuseppe Riccardi, and Fabio Pianesi. 2014. The workshop on computational personality recognition 2014. In *Proceedings of the ACM International Conference on Multimedia*, pages 1245–1246. ACM.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT), Lake Tahoe, CA*, pages 2–17.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL 1996)*, pages 310–318.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in statistical machine translation. In *Proceedings of ACL*.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Seattle, Washington, USA, October. Association for Computational Linguistics.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15. Springer-Verlag.

George F. Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of EMNLP*.

Alastair J. Gill, Carsten Brockmann, and Jon Oberlander. 2012. Perceptions of alignment and personality in generated dialogue. In *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, pages 40–48. Association for Computational Linguistics.

G. H. Golub and C. Reinsch. 1970. Singular value decomposition and least squares solutions. *Journal of Numerical Mathematics*, 14:403–420.

Li Gong, Aurélien Max, and François Yvon. 2012. Towards contextual adaptation for any-text translation. In *Proceedings of IWSLT*.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 421–432, Jeju Island, Korea.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Susan C. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, September.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *53rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL Demo and Poster Sessions*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of EMNLP-CoNLL*.

Matthias R. Mehl, Samuel D. Gosling, and James W. Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862–877, May.

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 and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal. Association for Computational Linguistics.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

Scott Nowson and Alastair J. Gill. 2014. Look! Who's Talking? Projection of Extraversion Across Different Social Contexts. In *Proceedings of WCPR14, Workshop on Computational Personality Recognition at ACMM (22nd ACM International Conference on Multimedia)*.

Scott Nowson, Julien Perez, Caroline Brun, Shachar Mirkin, and Claude Roux. 2015. XRCE Personal Language Analytics Engine for Multilingual Author Profiling. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September.

Scott Nowson. 2006. *The Language of Weblogs: A study of genre and individual differences.* Ph.D. thesis, University of Edinburgh.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of COLING/ACL-06: 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, July.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL 2003)*, ACL '03, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, Pennsylvania, USA.

James W Pennebaker, Kate G Niederhoffer, and Matthias R Mehl. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577, January.

Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212, February.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September.

Robert Schapire. 1990. The strength of weak learnability. *Journal of Machine Learning Research*, 5.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings Int. Conf. on Spoken Language Processing (INTERSPEECH 2002)*, pages 257–286.

Deborah Tannen. 1990. *You Just Don't Understand: Women and Men in Conversation*. Harper Collins, New York.

Marko Tkalčič, Berardina De Carolis, Marco de Gemmis, Ante Odić, and Andrej Košir. 2014. Preface: Empire 2014. In *Proceedings of the 2nd Workshop Emotions and Personality in Personalized Services (EMPIRE 2014)*. CEUR-WS.org, July.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *EMNLP*, pages 1815–1827. ACL.