Controlling Politeness in Neural Machine Translation via Side Constraints

Rico Sennrich and **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{rico.sennrich,a.birch}@ed.ac.uk, bhaddow@inf.ed.ac.uk

Abstract

Many languages use honorifics to express politeness, social distance, or the relative social status between the speaker and their addressee(s). In machine translation from a language without honorifics such as English, it is difficult to predict the appropriate honorific, but users may want to control the level of politeness in the output. In this paper, we perform a pilot study to control honorifics in neural machine translation (NMT) via *side constraints*, focusing on English→German. We show that by marking up the (English) source side of the training data with a feature that encodes the use of honorifics on the (German) target side, we can control the honorifics produced at test time. Experiments show that the choice of honorifics has a big impact on translation quality as measured by BLEU, and oracle experiments show that substantial improvements are possible by constraining the translation to the desired level of politeness.

1 Introduction

Many languages use honorifics to express politeness, social distance, or the relative social status between the speaker and their addressee(s). A widespread instance is the grammatical T-V distinction (Brown and Gilman, 1960), distinguishing between the familiar (Latin ***Tu***) and the polite (Latin ***Vos***) second person pronoun. In machine translation from a language without honorifics such as English, it is difficult to predict the appropriate honorific, but users may want to control the level of politeness in the output.

We propose a simple and effective method for including target-side T-V annotation in the training of a neural machine translation (NMT) system, which allows us to control the level of politeness at test time through what we call *side constraints*. It can be applied for translation between languages where the T-V distinction is missing from the source, or where the distribution differs. For instance, both Swedish and French make the T-V distinction, but reciprocal use of T pronouns is more widespread in Swedish than in French (Schüpbach et al., 2006). Hence, the Swedish form is not a reliable signal for the appropriate form in the French translation (or vice-versa).

Our basic approach of using side constraints to control target-side features that may be missing from the source, or are unreliable because of a category mismatch, is not limited to the T-V distinction, but could be applied to various linguistic features. This includes grammatical features such as tense and the number/gender of discourse participants, and more generally, features such as dialect and register choice.

This paper has the following contributions:

- we describe rules to automatically annotate the T-V distinction in German text.
- we describe how to use target-side T-V annotation in NMT training to control the level of politeness at test time via side constraints.
- we perform oracle experiments to demonstrate the impact of controlling politeness in NMT.

2 Background: Neural Machine Translation

Attentional neural machine translation (Bahdanau et al., 2015) is the current state of the art for

The research presented in this publication was conducted in cooperation with Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland.

Proceedings of NAACL-HLT 2016, pages 35–40,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

English→German (Jean et al., 2015b; Luong and Manning, 2015). We follow the neural machine translation architecture by Bahdanau et al. (2015), which we will briefly summarize here. However, our approach is not specific to this architecture.

The neural machine translation system is implemented as an attentional encoder-decoder network. The encoder is a bidirectional neural network with gated recurrent units (Cho et al., 2014) that reads an input sequence $x = (x_1, ..., x_m)$ and calculates a forward sequence of hidden states $(\overrightarrow{h_1}, ..., \overrightarrow{h_m})$, and a backward sequence $(\overleftarrow{h_1}, ..., \overleftarrow{h_m})$. The hidden states $\overrightarrow{h_j}$ and $\overleftarrow{h_j}$ are concatenated to obtain the annotation vector h_j.

The decoder is a recurrent neural network that predicts a target sequence $y = (y_1, ..., y_n)$. Each word y_i is predicted based on a recurrent hidden state s_i, the previously predicted word y_{i-1}, and a context vector c_i. c_i is computed as a weighted sum of the annotations h_j. The weight of each annotation h_j is computed through an *alignment model* α_{ij}, which models the probability that y_i is aligned to x_j. The alignment model is a single-layer feedforward neural network that is learned jointly with the rest of the network through backpropagation.

A detailed description can be found in (Bahdanau et al., 2015). Training is performed on a parallel corpus with stochastic gradient descent. For translation, a beam search with small beam size is employed.

3 NMT with Side Constraints

We are interested in machine translation for language pairs where politeness is not grammatically marked in the source text, but should be predicted in the target text. The basic idea is to provide the neural network with additional input features that mark *side constraints* such as politeness.

At training time, the correct feature is extracted from the sentence pair as described in the following section. At test time, we assume that the side constraint is provided by a user who selects the desired level of politeness of the translation.

We add side constraints as special tokens at the end of the source text, for instance <*T*> or <*V*>. The attentional encoder-decoder framework is then able to learn to pay attention to the side constraints. One could envision alternative architectures to incorporate side constraints, e.g. directly connecting them to all decoder hidden states, bypassing the attention model, or connecting them to the output layer (Mikolov and Zweig, 2012). Our approach is simple and applicable to a wide range of NMT architectures and our experiments suggest that the incorporation of the side constraint as an extra source token is very effective.

4 Automatic Training Set Annotation

Our approach relies on annotating politeness in the training set to obtain the politeness feature which we discussed previously. We choose a sentence-level annotation because a target-side honorific may have no word-level correspondence in the source. We will discuss the annotation of German as an example, but our method could be applied to other languages, such as Japanese (Nariyama et al., 2005).

German has distinct pronoun forms for informal and polite address, as shown in Table 1. A further difference between informal and polite speech are imperative verbs, and the original imperative forms are considered informal. The polite alternative is to use 3rd person plural forms with subject in position 2:

- *Ruf mich zurück.* (informal)
  (Call me back.)

- *Rufen Sie mich zurück.* (polite)
  (Call you me back.)

We automatically annotate politeness on a sentence level with rules based on a morphosyntactic annotation by ParZu (Sennrich et al., 2013). Sentences containing imperative verbs are labelled informal. Sentences containing an informal or polite pronoun from Table 1 are labelled with the corresponding class.

Some pronouns are ambiguous. Polite pronouns are distinguished from (neutral) 3rd person plural forms by their capitalization, and are ambiguous in sentence-initial position. In sentence-initial position, we consider them polite pronouns if the English source side contains the pronoun *you(r)*. For *Ihr* and *ihr*, we use the morphological annotation by ParZu to distinguish between the informal 2nd person plural nominative, the (neutral) 3rd person singular dative, and the possessive; for possessive pronouns, we

category	informal		polite
	sg.	pl.	sg./pl.
nominative	du	ihr	Sie
genitive	deiner	euer	Ihrer
dative	dir	euch	Ihnen
accusative	dich	euch	Sie
possessive (base form)	dein	euer	Ihr

Table 1: German address pronouns.

distinguish between polite forms and (neutral) 3rd person forms by their capitalization.

If a sentence matches rules for both classes, we label it as informal – we found that our lowest-precision rule is the annotation of sentence-initial *Sie*. All sentences without a match are considered neutral.

5 Evaluation

Our empirical research questions are as follows:

- can we control the production of honorifics in neural machine translation via side constraints?
- how important is the T-V distinction for translation quality (as measured by BLEU)?

5.1 Data and Methods

We perform English→German experiments on OpenSubtitles (Tiedemann, 2012)[1], a parallel corpus of movie subtitles. Machine translation is commonly used in the professional translation of movie subtitles in a post-editing workflow, and politeness is considered an open problem for subtitle translation (Etchegoyhen et al., 2014). We use OpenSubtitles2012 as training corpus, and random samples from OpenSubtitles2013 for testing. The training corpus consists of of 5.58 million sentence pairs, out of which we label 0.48 million sentence pairs as polite, and 1.09 million as informal.

We train an attentional encoder-decoder NMT system using Groundhog[2] (Bahdanau et al., 2015; Jean et al., 2015a). We follow the settings and training procedure described by Sennrich et al. (2015), using BPE to represent the texts with a fixed vocabulary of subword units (vocabulary size 90000).

The training set is annotated as described in section 4, and the source side is marked with the politeness feature as described in section 3. Note that there are only two values for the politeness feature, and neutral sentences are left unmarked. This is to allow users to select a politeness level for the whole document, without having to predict which translations should contain an address pronoun. Instead, we want the NMT model to ignore side constraints when they are irrelevant.

To ensure that the NMT model does not overly rely on the side constraints, and that performance does not degrade when no side constraint is provided at test time, only a subset of the labelled training instances are marked with a politeness feature at training time. We set the probability that a labelled training instance is marked, α, to 0.5 in our experiments. To ensure that the NMT model learns to ignore side constraints when they are irrelevant, and does not overproduce address pronouns when side constraints are active, we also mark neutral sentences with a random politeness feature with probability α. Keeping the mark-up probability α constant for all sentences in the training set prevents the introduction of unwanted biases. We re-mark the training set for each epoch of training. In preliminary experiments, we found no degradation in baseline performance when politeness features were included in this way during training.

The model is trained for approximately 9 epochs (7 days). At test time, all results are obtained with the same model, and the only variable is the side constraint used to control the production of honorifics. We test translation without side constraint, and translations that are constrained to be polite or informal. In an oracle experiment, we use the politeness label of the reference to determine the side constraint. This simulates a setting in which a user controls the desired politeness.

5.2 Results

Our first test set is a random sample of 2000 sentences from OpenSubtitles2013 where the English source contains a 2nd person pronoun. Results are shown in Table 2. Side constraints very effectively control whether the NMT system produces polite or informal output. Translations constrained to be polite are overwhelmingly labelled polite or neutral

[1] http://www.opensubtitles.org

[2] github.com/sebastien-j/LV_groundhog

side constraint	output label			BLEU
	neutral	polite	informal	
(reference)	429	524	1047	-
none	178	351	1471	20.7
polite	208	1728	64	17.9
informal	141	28	1831	20.2
oracle	161	567	1272	23.9

Table 2: Politeness and translation quality on test set of 2000 sentences from OpenSubtitles2013 that contain second person pronoun *you(r(s(elf)))* in English source text.

source	Give me the telephone!
reference	Gib mir das Telefon! [T]
none	Gib mir das Telefon! [T]
polite	Geben Sie mir das Telefon! [V]
informal	Gib mir das Telefon! [T]
source	Are you kidding?
reference	Das ist doch ein Witz! [N]
	(this is a joke!)
none	Machst du Witze? [T]
polite	Machen Sie Witze? [V]
informal	Machst du Witze? [T]
source	You foolish boy.
reference	Du dummer Junge. [T]
none	Du dummer Junge. [T]
polite	Du dummer Junge. [T]
informal	Du dummer Junge. [T]

Table 3: Translation examples with different side constraints. Translations marked as neutral [N], informal [T] or polite [V].

side constraint	output label			BLEU
	neutral	polite	informal	
(reference)	1406	189	405	-
none	1385	125	490	22.6
polite	1386	576	38	21.7
informal	1365	11	624	22.5
oracle	1374	185	441	24.0

Table 4: Politeness and translation quality on test set of 2000 random sentences from OpenSubtitles2013.

by our automatic target-side annotation (96%), and analogously, translations constrained to be informal are almost exclusively informal or neutral (98%).

We also see that BLEU is strongly affected by the choice. An oracle experiment in which the side constraint of every sentence is informed by the reference obtains an improvement of 3.2 BLEU over the baseline (20.7→23.9).

We note that the reference has a higher proportion of German sentences labelled neutral than the NMT systems. A close inspection shows that this is due to sentence alignment errors in OpenSubtitles, free translations as shown in Table 3, and sentences where *you* is generic and translated by the impersonal pronoun *man* in the reference.

The side constraints are only soft constraints, and are occasionally overridden by the NMT system. These cases tend to be sentences where the source text provides strong politeness clues, like the sentence *You foolish boy*. Neither the address *boy* nor the attribute *foolish* are likely in polite speech, and the sentence is translated with a T pronoun, regardless of the side constraint.

While Table 2 only contains sentences with an address pronoun in the source text, Table 4 represents a random sample. There are fewer address pronouns in the random sample, and thus more neutral sentences, but side constraints remain effective. This experiment also shows that we do not overproduce address pronouns when side constraints are provided, which validates our strategy of including side constraints with a constant probability α at training time.

The automatic evaluation with BLEU indicates that the T-V distinction is relevant for translation. We expect that the actual relevance for humans depends on the task. For gisting, we expect the T-V distinction to have little effect on comprehensibility. For professional translation that uses MT with post-editing, producing the desired honorifics is likely to improve post-editing speed and satisfaction. In an evaluation of MT for subtitle translation, Etchegoyhen et al. (2014) highlight the production of the appropriate T-V form as “a limitation of MT technology” that was “often frustrat[ing]” to post-editors.

6 Related Work

Faruqui and Pado (2012) have used a bilingual English–German corpus to automatically annotate the T-V distinction, and train a classifier to predict the address from monolingual English text. Applying a source-side classifier is potential future work, although we note that the baseline encoder–decoder NMT system already has some disambiguating power. Our T-V classification is more comprehensive, including more pronoun forms and imperative verbs.

Previous research on neural language models has proposed including various types of extra information, such as topic, genre or document context (Mikolov and Zweig, 2012; Aransa et al., 2015; Ji et al., 2015; Wang and Cho, 2015). Our method is somewhat similar, with the main novel idea being that we can target specific phenomena, such as honorifics, via an automatic annotation of the target side of a parallel corpus. On the modelling side, our method is slightly different in that we pass the extra information to the encoder of an encoder–decoder network, rather than the (decoder) hidden layer or output layer. We found this to be very effective, but trying different architectures is potential future work.

In rule-based machine translation, user options to control the level of politeness have been proposed in the 90s (Mima et al., 1997), and were adopted by commercial systems (SYSTRAN, 2004, 26). To our knowledge, controlling the level of politeness has not been explicitly addressed in statistical machine translation. While one could use data selection or weighting to control the honorifics produced by SMT, NMT allows us to very elegantly support multiple levels of politeness with a single model.

7 Conclusion

Machine translation should not only produce semantically accurate translations, but should also consider pragmatic aspects, such as producing socially appropriate forms of address. We show that by annotating the T-V distinction in the target text, and integrating the annotation as an additional input during training of a neural translation model, we can apply side constraints at test time to control the production of honorifics in NMT.

We currently assume that the desired level of politeness is specified by the user. Future work could aim to automatically predict it from the English source text based on textual features such as titles and names, or meta-textual information about the discourse participants.

While this paper focuses on controlling politeness, side constraints could be applied to a wide range of phenomena. It is a general problem in translation that, depending on the language pair, the translator needs to specify features in the target text that cannot be predicted from the source text. Apart from from the T-V distinction, this includes grammatical features such as clusivity, tense, and gender and number of the discourse participants, and more generally, features such as the desired dialect (e.g. when translating into Arabic) and text register. Side constraints can be applied to control these features. All that is required is that the feature can be annotated reliably, either using target-side information or metatextual information, at training time.

Acknowledgments

The research presented in this publication was conducted in cooperation with Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 644402 (HimL).

References

Walid Aransa, Holger Schwenk, and Loïc Barrault. 2015. Improving Continuous Space Language Models using Auxiliary Features. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 151–158, Da Nang, Vietnam.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Roger Brown and A. Gilman. 1960. The pronouns of power and solidarity. In T. Sebeok, editor, *Style in Language*. The M.I.T. Press, Cambridge, MA.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maucec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Manaal Faruqui and Sebastian Pado. 2012. Towards a model of formal and informal address in English. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–633, Avignon, France. Association for Computational Linguistics.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015a. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015b. Montreal Neural Machine Translation Systems for WMT'15 . In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2015. Document Context Language Models. *ArXiv e-prints*, November.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation 2015*, Da Nang, Vietnam.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239, Miami, FL, USA.

Hideki Mima, Osamu Furuse, and Hitoshi Iida. 1997. Improving Performance of Transfer-driven Machine Translation with Extra-linguistic Information from Context, Situation and Environment. In *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence - Volume 2*, IJCAI'97, pages 983–988, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shigeko Nariyama, Hiromi Nakaiwa, and Melanie Siegel. 2005. Annotating Honorifics Denoting Social Ranking of Referents. In *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (LINC-2005)*.

Doris Schüpbach, John Hajek, Jane Warren, Michael Clyne, Heinz-L. Kretzenbacher, and Catrin Norrby. 2006. A cross-linguistic comparison of address pronoun use in four European languages: Intralingual and interlingual dimensions . In *Annual Meeting of the Australian Linguistic Society*, Brisbane, Australia.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*, pages 601–609, Hissar, Bulgaria.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *CoRR*, abs/1508.07909.

SYSTRAN, 2004. *SYSTRAN 5.0 User Guide*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Tian Wang and Kyunghyun Cho. 2015. Larger-Context Language Modelling. *ArXiv e-prints*, November.