A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output

Xing Niu
Dept. of Computer Science
University of Maryland
College Park
xingniu@cs.umd.edu

Marianna Martindale
iSchool
University of Maryland
College Park
mmartind@umd.edu

Marine Carpuat
Dept. of Computer Science
University of Maryland
College Park
marine@cs.umd.edu

Abstract

Stylistic variations of language, such as formality, carry speakers' intention beyond literal meaning and should be conveyed adequately in translation. We propose to use lexical formality models to control the formality level of machine translation output. We demonstrate the effectiveness of our approach in empirical evaluations, as measured by automatic metrics and human assessments.

1 Introduction

Automatically analyzing and generating natural language requires capturing not only what is said, but also how to say it. Consider the sentences "anybody hurt?" and "is someone wounded?". The first one is less formal than the second one, and carries information beyond its literal meaning, such as the situation in which it might be used. Such differences in formality have been identified as an important dimension of style (Trudgill, 1992) or tone (Halliday, 1978) variation.

In this paper, we build on prior computational work that has focused on analyzing formality of texts (Lahiri and Lu, 2011; Brooke and Hirst, 2013; Pavlick and Nenkova, 2015; Pavlick and Tetreault, 2016) with a different aim: modeling formality for the purpose of controlling style in applications that generate language, with a focus on machine translation. Human translators translate a document for a specific audience (Nida and Taber Charles, 1969), and often ask what is the expected tone of the content when taking a new translation job. We design a machine translation system that operates under similar conditions and explicitly takes an expected level of formality as input. While ultimately we would like systems to preserve the formality of the source, this is a challenging task that requires not only automatically inferring the formality of the source, but also understanding how formality differs across languages and cultures. As a first step, we therefore limit our study to the scenario where the expected output formality is given to the MT system as an additional input.

We first select a formality model providing the most accurate scores on intrinsic formality datasets. We compare existing lexical formality models and novel variants based on inducing formality dimensions or subspaces in vector space models. We then turn to machine translation and show that a lexical formality model can have a positive impact when used to control the formality of machine translation output. When the expected formality matches the reference, we obtain improvement of translation quality evaluated by automatic metrics (BLEU). A human assessment also verified the effectiveness of our proposed system in generating translations at diverse levels of formality.

2 Formality-Sensitive MT

Our goal is to provide systems with the ability to generate language across a range of formality style. We propose a **Formality-Sensitive Machine Translation (FSMT)** scenario where the system takes two inputs: (1) text in the source language to be translated, and (2) a desired formality level capturing the intended audience of the translation. We propose to implement it as n-best re-ranking within a standard phrase-based MT architecture. Unlike domain adaptation approaches, which aim to produce domain-specific or potentially formality-specific systems, our goal is to obtain a single system trained on diverse data which can adaptively produce output for a range of styles.

We therefore introduce a formality-scoring fea-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2814–2819
Copenhagen, Denmark, September 7–11, 2017.

ture for re-ranking. For each translation hypothesis h, given the formality level ℓ as a parameter:

$$f(h;\ell) = |\,\text{Formality}(h) - \ell|$$

where $\text{Formality}(h)$ is the sentence-level formality score for h. $f(h;\ell)$, along with standard model features, is fed into a standard re-ranking model. When training the re-ranking model, the parameter ℓ is set to the actual formality score of the reference translation for each instance. At test time, ℓ is provided by the user. The re-scoring weights help promote candidate sentences whose formality scores approach the expected level.

3 Formality Modeling

The FSMT system requires quantifying the formality level of a sentence. Following prior work, we define sentence-level formality based on lexical formality scores (Brooke et al., 2010; Pavlick and Nenkova, 2015). We conduct an empirical comparison of existing techniques that can be adapted as lexical formality models, and introduce a sentence-level formality scheme based on weighted average.

3.1 Lexical Formality

State-of-the-art lexical formality models (Brooke et al., 2010; Brooke and Hirst, 2014) are based on vector space models of word meaning, and a set of pre-selected seed words that are representative of formal and informal language.

`SimDiff` Brooke et al. (2010) proposed to score the formality of a word w by comparing its meaning to that of seed words of known formality using cosine similarity. Intuitively, w is more likely formal if it is semantically closer to formal seed words than to informal seed words. Formally, given a formal word set S_f and an informal word set S_i, `SimDiff` scores a word w by

$$\text{score}(w) = \frac{1}{|S_f|}\sum_{v\in S_f}\cos(e_w, e_v) - \frac{1}{|S_i|}\sum_{v\in S_i}\cos(e_w, e_v)$$

Turning this difference into a formality score requires further manipulation. A neutral word r has to be manually selected to anchor the midpoint of the formality score range. In other words, the final formality score for r is enforced to be zero:

$$\text{Formality}(w) = \frac{\text{score}(w) - \text{score}(r)}{\text{normalizer}(w, r)}$$

The neutral word is typically selected from function words. We select "at" because it appears in nearly every document and appears with nearly equivalent probabilities in formal/informal corpora. Finally, a normalizer which is maximized among the whole vocabulary ensures that scores cover the entire $[-1, 1]$ range.

Instead of using cosine diff as the score function $\text{score}(w)$, other standard techniques can be also applied under this framework.

`SVM` As an alternative to the model proposed by Brooke and Hirst (2014), we propose to train an Support Vector Machine (`SVM`) model to find a hyperplane that separates formal and informal words and define the score function as the distance to the hyperplane.

Formality Subspace Another category of methods consists in identifying a subspace that captures formality within the original vector space. Lexical scores can then simply be obtained by projecting word representations onto the formality subspace. One example is training a Principal Component Analysis (**`PCA`**) model on word representations of all seeds. This method is based on the assumption that representative formal/informal words principally vary along the direction of formality. Alternatively, inspired by **DENSIFIER** (Rothe et al., 2016), we can learn a subspace that aims at separating words in S_f vs. words in S_i and grouping words in the same set.

3.2 From Word to Sentence Formality

While previous work scored sentence by averaging word scores (Brooke and Hirst, 2014; Pavlick and Nenkova, 2015), we propose a weighted average scheme for word sequences W to downgrade the formality contribution of neutral words:

$$\text{Formality}(W) = \frac{\sum_{w_i\in W} |\,\text{Formality}(w_i)| \cdot \text{Formality}(w_i)}{\sum_{w_i\in W} |\,\text{Formality}(w_i)|}$$

3.3 Evaluation

Before evaluating our FSMT framework, we evaluate the formality models at the sentence level. Lahiri (2015) and Pavlick and Tetreault (2016) collected 5-way human scores for 11,263 sentences in the genres of blog, email, answers and news. Following Pavlick and Tetreault (2016), we averaged human scores for each sentence as the

gold standard. As in prior work, the score quality was evaluated by the Spearman correlation.

A large mixed-topic corpus is required to train vector space models. As suggested by Brooke et al. (2010), we used the ICWSM 2009 Spinn3r dataset (English tier-1) which consists of about 1.6 billion words (Burton et al., 2009). We also compared the term-document association model Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and the term-term association model word2vec (W2V) (Mikolov et al., 2013). We used the same 105 formal seeds and 138 informal seeds as Brooke et al. (2010).

Followed Brooke et al. (2010), to achieve best performance, we used a small dimensionality (10) for training LSA and word2vec. In practice, we normalized the LSA word vectors to make them have unit length for SVM and PCA, but did not applied it to word2vec. This suggests that the magnitude of LSA word vectors is harmful for formality modeling.

We also compared formality models based on word representations to a baseline that relies on unigram models to compare word statistics in corpora representative of formal vs. informal language (Pavlick and Nenkova, 2015). This method requires language examples of diverse formality. Conversational transcripts are generally considered as casual text, so we concatenated corpora such as Fisher (Cieri et al., 2004), Switchboard (Godfrey et al., 1992), SBCSAE (Bois et al., 2000-2005), CallHome[1], CallFriend[2], BOLT SMS/Chat (Song et al., 2014) and NPS Chatroom (Forsythand and Martell, 2007). As the formal counterpart, we extracted comparable size of text from Europarl (Koehn, 2005). This results in 30 Million tokens of formal corpora (1.1M segments) and 29 Million tokens of informal corpora (2.7M segments).

Table 1 shows that all models based on the vector space achieve similar performance in terms of Spearman's ρ (except SVM-W2V which yields lower performance). The baseline method based on unigram models was outperformed by 0.1+ point. So we select DENSIFIER-LSA as a representative for our FSMT system.

[1] https://catalog.ldc.upenn.edu/LDC97S42

[2] https://talkbank.org/access/CABank/CallFriend/

	LSA	W2V
SimDiff	0.660	0.654
SVM	0.657	0.585
PCA	0.656	0.663
DENSIFIER	0.664	0.644
baseline	0.540	

Table 1: Sentence-level formality quantifying evaluation (Spearman's ρ) among different models with different vector spaces.

4 Evaluation of the FSMT System

Set-up We evaluate this approach on a French to English translation task. Two parallel French-English corpora are used: (1) MultiUN (Eisele and Chen, 2010), which is extracted from the United Nations website, and can be considered to be formal text; (2) OpenSubtitles2016 (Lison and Tiedemann, 2016), which is extracted from movie and TV subtitles, covers a wider spectrum of styles, but overall tends to be informal since it primarily contains conversations. Each parallel corpus was split into a training set (100M English tokens), a tuning set (2.5K segments) and a test set (5K segments). Two corpora are then concatenated, such that training, tuning and test sets all contained a diversity of styles.

Moses (Koehn et al., 2007) is used to build our phrase-based MT system. We followed the standard training pipeline with default parameters.[3] Word alignments were generated using fast_align (Dyer et al., 2013), and symmetrized using the *grow-diag-final-and* heuristic. We used 4-gram language models, trained using KenLM (Heafield, 2011). Model weights were tuned using batch MIRA (Cherry and Foster, 2012).

We used constant size n=1000 for n-best lists in all experiments. The re-ranking is a log-linear model trained using batch MIRA. [4] We report results averaged over 5 random tuning re-starts to compensate for tuning noise (Clark et al., 2011).

FSMT In order to evaluate the impact of different input formality (e.g. low/neutral/high) on translation quality, ideally, we would like to have three human reference translations with different

[3] http://www.statmt.org/moses/?n=Moses.Baseline

[4] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/nbest-rescore

Desired formality	Informal test set	Neutral test set	Formal test set
None (baseline)	39.74	40.17	**47.97**
low	**40.27**	39.65	47.76
neutral	38.70	**40.46**	47.84
high	37.58	39.53	**47.97**

Table 2: Translation quality (BLEU scores) on informal/neutral/formal sentence sets given different desired formality levels (-0.4, 0.0, 0.4). Best results with statistical significance are highlighted.

formality for each source sentence. Since such references are not available, we construct three sets of test data where instances are divided according to the formality level of the available reference translation. The formality distribution in the tuning set shows that 97% reference translations fall into the range of $[-0.6, 0.6]$. We therefore set three formality bins – informal $[-1, -0.2)$, neutral formality $[-0.2, 0.2]$, and formal $(0.2, 1]$ – and split the test set into these bins. We use DENSIFIER-LSA and training setting described above to translate the entire test set three times, with three different formality levels: low (-0.4), neutral (0) and high (0.4).

4.1 Automatic Evaluation

We first report standard automatic evaluation results using the BLEU score to compare FSMT output given different desired formality level on each bins (See Table 2).

The best BLEU scores for each formality level are obtained when the level of formality given as input to the MT system matches the nature of the text being translated, as can be seen in the scores along the diagonal in Table 2. Comparing with the baseline system, which produces the top translation from each n-best list, translation quality improves by +0.5 BLEU on informal text, +0.3 BLEU on neutral text, and remains constant on formal text. The impact increases with the distance to formal language increases. This can be explained by the fact that more formal sentences tend to be longer, and the impact of alternate lexical choice for a small number of words per sentence is smaller in longer sentences. In addition, the formal sentences are mostly drawn from UN data which is sufficiently different from the other genres in the heterogeneous training corpus that the informal examples do not affect baseline performance on formal data.

4.2 Human Assessment

Automatic evaluation is limited to comparing output to a single reference: lower BLEU scores conflate translation errors and stylistic mismatch. Therefore, we conduct a human study of the formality vs. the quality.

We conducted a manual evaluation of the output of our FSMT system taking low/high formality levels (-0.4/0.4) as parameters. 42 translation pairs were randomly selected and were annotated by 15 volunteers. For each pair of segments, the volunteers were asked to select the segment that would be more appropriate in a formal setting (e.g., a job interview) than in a casual setting (e.g., chatting with friends). A default option of "neither of them is more formal or hard to say" was also available.

By majority voting, 20 pairs were annotated as "N", indicating the two translations has no distinctions w.r.t. formality. For example, "A: how can they do this" vs. "B: how can they do that". Given that the translations were restricted to the n-best list, not all sentences could be translated into stylistically different language.

Of the remaining 21 pairs where annotators judged one output more formal than the other, in all but one case the translation produced by our FSMT system with high formality level parameter was judged to be more formal. Overall this indicates that our formality scoring and ranking procedure are effective.

To determine whether re-ranking based on formality might have a detrimental effect on quality, we also had annotators rate the fluency and adequacy of the segments. Inspired by Graham et al. (2013), annotators were first asked to assess fluency without a reference and separately adequacy with a reference. Both assessments used a sliding scale. Each segment was evaluated by an average of 7 annotators. After rescaling the ratings into the $[0, 1]$ range, we observed a 0.75 level of fluency for informal translations and 0.70 for formal ones. This slight difference fits our expectation that more casual language may feel more fluent while more formal language may feel more stilted. The adequacy ratings were 0.65 and 0.64 for informal and translations respectively, indicating that adjusting the level of formality had minimal effect on the adequacy of the result.

Some examples are listed in Table 3. Occa-

ℓ	Examples	Comments
-0.4	... and then he **ran away** .	–
0.4	... and then he **escaped** .	*annotated as more formal*
-0.4	**anybody hurt** ?	–
0.4	is **someone wounded** ?	*annotated as more formal*
-0.4	he **shot himself** in the middle of it .	–
0.4	he **committed suicide** in the middle of it .	*annotated as more formal*
-0.4	to move **things** forward .	–
0.4	**in order** to move **the process** forward.	*annotated as more formal*
-0.4	how do you do ?	*annotated as more formal*
0.4	how are you?	–
-0.4	oh , val , you should get the phone .	*missing words*
0.4	oh , val , you should have the phone (of pete) .	–
-0.4	**i believe** you've solved the case , lieutenant .	*additive words*
0.4	you solved the case , lieutenant .	–
REF	right by checkout .	
-0.4	right next to the **body** .	*incorrect word choice*
0.4	right next to the **fund** .	*incorrect word choice*

Table 3: Examples of variant translations to the same French source segment using low/high output formality levels (-0.4/0.4) as parameters. In general the variations lie on the direction of formality as expected, but occasionally translation errors occur.

sionally, the n-best list had no translation hypotheses with diverse formality, so the FSMT system dropped necessary words, appended inessential words, or selected improper or even incorrect words to fit the target formality level. In the case of 'how do you do', the translation that was meant to be more casual was rated more formal. Because the system measures formality on the lexical level, it was not able to recognize this idiomatically formal phrase made up of words that are not inherently formal. Despite these issues, most of the output were formality-variant translations of the same French source segment, as expected.

5 Conclusion

We presented a framework for formality-sensitive machine translation, where a system produces translations at a desired formality level. Our evaluation shows the effectiveness of this system in controlling language formality without loss in translation quality.

References

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa barbara corpus of spoken american english, parts 1-4.

Julian Brooke and Graeme Hirst. 2013. Hybrid models for lexical acquisition of correlated styles. In *IJCNLP*, pages 82–90. Asian Federation of Natural Language Processing / ACL.

Julian Brooke and Graeme Hirst. 2014. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *COLING*, pages 2172–2183. ACL.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *COLING (Posters)*, pages 90–98. Chinese Information Processing Society of China.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA*.

Colin Cherry and George F. Foster. 2012. Batch tuning strategies for statistical machine translation. In *HLT-NAACL*, pages 427–436. The Association for Computational Linguistics.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*. European Language Resources Association.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL (Short Papers)*, pages 176–181. The Association for Computer Linguistics.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *HLT-NAACL*, pages 644–648. The Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*. European Language Resources Association.

Eric N. Forsythand and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *ICSC*, pages 19–26. IEEE Computer Society.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *LAW@ACL*, pages 33–41. The Association for Computer Linguistics.

Michael AK Halliday. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

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 *ACL*. The Association for Computational Linguistics.

Shibamouli Lahiri. 2015. SQUINKY! A corpus of sentence-level formality, informativeness, and implicature. *CoRR*, abs/1506.02306.

Shibamouli Lahiri and Xiaofei Lu. 2011. Inter-rater agreement on sentence formality. *CoRR*, abs/1109.0069.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC*. European Language Resources Association (ELRA).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

EugeneA. Nida and R Taber Charles. 1969. The theory and practice of translation. In *Helps for Translators*, volume 8. United Bible Societies.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *HLT-NAACL*, pages 218–224. The Association for Computational Linguistics.

Ellie Pavlick and Joel R. Tetreault. 2016. An empirical analysis of formality in online communication. *TACL*, 4:61–74.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *HLT-NAACL*, pages 767–777. The Association for Computational Linguistics.

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, and Ann Sawyer. 2014. Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *LREC*, pages 1699–1704. European Language Resources Association (ELRA).

Peter Trudgill. 1992. *Introducing Language and Society*. Penguin, London.