Associative Texture is Lost in Translation

Beata Beigman Klebanov and Michael Flor
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541
{bbeigmanklebanov,mflor}@ets.org

Abstract

We present a suggestive finding regarding the loss of associative texture in the process of machine translation, using comparisons between (a) original and back-translated texts, (b) reference and system translations, and (c) better and worse MT systems. We represent the amount of association in a text using **word association profile** – a distribution of pointwise mutual information between all pairs of content word types in a text. We use the average of the distribution, which we term **lexical tightness**, as a single measure of the amount of association in a text. We show that the lexical tightness of human-composed texts is higher than that of the machine translated materials; human references are tighter than machine translations, and better MT systems produce lexically tighter translations. While the phenomenon of the loss of associative texture has been theoretically predicted by translation scholars, we present a measure capable of quantifying the extent of this phenomenon.

1 Introduction

While most current approaches to machine translation concentrate on single sentences, there is emerging interest in phenomena that go beyond a single sentence and pertain to the whole text being translated. For example, Wong and Kit (2012) demonstrated that repetition of content words is a predictor of translation quality, with poorer translations failing to repeat words appropriately. Gong et al. (2011) and Tiedemann (2010) present caching of translations from earlier sections of a document to facilitate the translation of its later sections.

In scholarship that deals with properties of human translation of literary texts, translation is often rendered as a process that tends to *deform* the original, and a number of particular aspects of deformation have been identified. Specifically, Berman (2000) discusses the problem of *quantitative impoverishment* thus:

> This refers to a lexical loss. Every work in prose presents a certain *proliferation* of signifiers and signifying chains. Great novelist prose is "abundant." These signifiers can be described as *unfixed*, especially as a signified may have a multiplicity of signifiers. For the signified *visage* (face) Arlt employs *semblante*, *rosto* and *cara* without justifying a particular choice in a particular sentence. The essential thing is that *visage* is marked as an important *reality* in his work by the use of three signifiers. The translation that does not respect this multiplicity renders the "visage" of an unrecognizable work. There is a loss, then, since the translation contains fewer signifiers than the original."[1]

While Berman's remarks refer to literary translation, recent work demonstrates its relevance for machine translation, showing that MT systems tend to under-use linguistic devices that are commonly used for repeated reference, such as superordinates or meronyms, although the pattern with synonyms and near-synonyms was not clear cut (Wong and Kit, 2012). Studying a complementary phenomenon of translation of same-lemma lexical items in the source document into a target language, Carpuat and Simard (2012) found that when MT systems produce different target language translations, they are stylistically, syntactically, or semantically inadequate in most cases

[1] italics in the original

Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), pages 27–32,
Sofia, Bulgaria, August 9, 2013. ©2013 Association for Computational Linguistics

(see upper panel of Table 5 therein), that is, diversifying the signifiers appropriately is a challenging task. For recent work on biasing SMT systems towards consistent translations of repeated words, see Ture et al. (2012) and Xiao et al. (2011).

Moving beyond single signifieds, or concepts, Berman faults translations for "the destruction of underlying networks of signification", whereby groups of related words are translated without preserving the relatedness in the target language. While these might be unavoidable in any translation, we show below that machine translation specifically indeed suffers from such a loss (section 3) and that machine translation suffers from it more than the human translations (section 4).

2 Methodology

We define $\mathbf{WAP_T}$ – a **word association profile** of a text T – as the distribution of PMI(x, y) for all pairs of content[2] word types (x, y) ∈T.[3] We estimate PMIs using same-paragraph co-occurrence counts from a large and diverse corpus of about 2.5 billion words: 2 billion words come from the Gigaword 2003 corpus (Graff and Cieri, 2003); an additional 500 million words come from an inhouse corpus containing popular science and fiction texts. We further define $\mathbf{LT_T}$ – the **lexical tightness** of a text T – as the average value of the word association profile. All pairs of words in T for which the corpus had no co-occurrence data are excluded from the calculations. We note that the database has very good coverage with respect to the datasets in sections 3-5, with 94%-96% of pairs on average having co-occurrence counts in the database. A more detailed exposition of the notion of a word association profile, including measurements on a number of corpora, can be found in Beigman Klebanov and Flor (2013).

Our prediction is that translated texts would be less lexically tight than originals, and that better translations – either human or machine – would be tighter than worse translations, incurring a smaller amount of association loss.

3 Experiment 1: Back-translation

For the experiment, we selected 20 editorials on the topic of baseball from the New York Times Annotated Corpus.[4] The selected articles had baseball annotated as their sole topic, and ranged from 250 to 750 words in length. We expect these articles to contain a large group of words that reflects vocabulary that is commonly used in discussing baseball and no other systematic subtopics. All articles were translated into French, Spanish, Arabic, and Swedish, and then translated back to English, using the Google automatic translation service. Our goal is to observe the effect of the two layers of translation (out of English and back) on the lexical tightness of the resulting texts.

Since baseball is not a topic that is commonly discussed in the European languages or in Arabic, this is a case where culturally foreign material needs to be rendered in a host (or target) language. This is exactly the kind of situation where we expect deformation to occur – the material is either altered so that is feels more "native" in the host language (domestication) or its foreigness is preserved (foreignization) in that the material lacks associative support in the host language (Venuti, 1995). In the first case, the translation might be associatively adequate *in the host language*, but, being altered, it would produce less culturally precise result when translated back into English. In the second case, the result of translating out of English might already be associatively impoverished *by the standards of the host language*.

The italicized phrases in the previous paragraph underscore the theoretical and practical difficulty in diagnozing domestication or foreignization in translating out of English – an associative model for each of the host languages will be needed, as well as some benchmark of the lexical tightness of native texts written on the given topic against which translations from English could be judged. While the technique of back-translation cannot identify the exact path of association loss – through domestication or foreignization – it can help *establish that association loss has occurred* in at least one or both of the translation processes involved, since the original native English version provides a natural benchmark against which the resulting back-translations can be measured.

To make the phenomenon of association loss more concrete, consider the following sentence:

Original Dave Magadan, the *hard-hitting rookie third baseman* groomed to replace Knight, has been hospitalized.

[2] We part-of-speech tag a text using OpenNLP tagger (http://opennlp.apache.org) and only take into account common and proper nouns, verbs, adjectives, and adverbs.

[3] PMI = Pointwise Mutual Information

[4] LDC2008T19 in LDC catalogue

Arabic Dave Magadan, the *stern rookie 3 baseman* groomed to replace Knight, is in the hospital.[5]

Spanish Dave Magadan, the *strong rookie third baseman* who managed to replace Knight, has been hospitalized.

French Dave Magadan, the *hitting third rookie player* prepared to replace Knight, was hospitalized.

Swedish Dave Magadan, *powerful rookie third baseman* groomed to replace Knight, has been hospitalized.

Observe the translations of the phrase "hard-hitting rookie third baseman." While substituting *strong* and *powerful* for *hard-hitting* might seem acceptable semantically, these terms are not associated with the other baseball terms in the text, whereas *hitting* is highly associated with them:[6] Table 1 shows PMI scores for each of *hitting, stern, strong, powerful* with the baseball terms *rookie* and *baseman*. The French translation got the *hitting*, but substituted the more generic term *player* instead of the baseball-specific *baseman*. As the bottom panel of Table 1 makes clear, while *player* is associated with other baseball terms, the associations are lower than those of *baseman*.

	rookie	baseman	hitting
hitting	3.54	5.29	
stern	0.35	-1.60	
strong	0.54	-0.08	
powerful	-0.62	-0.63	
player	3.95		2.73
baseman	5.11		5.29

Table 1: PMI associations of words introduced in back-translations with baseball terms *rookie*, *baseman*, and *hitting*.

Table 2 shows the average lexical tightness values across 20 texts for the original version as well as for the back translated versions. The original version is statistically significantly tighter than each of the back translated versions, using 4 applications of t-test for correlated samples, n=20, p<0.05 in each case.

Version	Av. LT	Std. LT	Min. LT	Max. LT
Original	**.953**	.092	.832	1.144
Via Arabic	.875	.093	.747	1.104
Via Spanish	.909	.081	.801	1.069
Via French	.912	.087	.786	1.123
Via Swedish	.931	.099	.796	1.131

Table 2: Average lexical tightness (Av. LT) for the original vs back translated versions, on 20 baseball texts from the New York Times. Standard deviation, minimum, and maximum values are also shown.

4 Experiment 2: Reference vs Machine Translation

We use a part of the dataset used in the NIST Open MT 2008 Evaluation.[7] Our set contains translations of 120 news and web articles from Arabic to English. For each document, there are 4 human reference translations and 17 machine translations by various systems that participated in the benchmark. Table 3 shows the average and standard deviation of lexical tightness values across the 120 texts for each of the four reference translations, each of the 17 MT systems, as well as an average across the four reference translations, and an average across the 17 MT systems. Each of the 17 MT systems is statistically significantly less tight than the average reference human translation (17 applications of the t-test for correlated samples, n=120, p<0.05); 12 of the 17 MT systems are statistically significantly less tight than the *least* tight human reference (reference translation #3) at p<0.05; the average system translation is statistically significantly less tight that the average human translation at p<0.05.

To exemplify a large gap in associative texture between reference and machine translations, consider the following extracts.[8] As the raw MT version (MT-raw) is barely readable, we provide a version where words are re-arranged for readability (MT-read), preserving most of the vocabulary. Since lexical tightness operates on content word types, adding or removing repetitions and function words does not impact the calculation, so we removed or inserted those for the sake of readability

[5] We corrected the syntax of all back-translations while preserving the content-word vocabulary choices.

[6] Our tokenizer splits words on hyphens, therefore examples are shown for *hitting* rather than for *hard-hitting*. The point still holds, since *hitting* is a baseball term on its own.

[7] LDC2010T01

[8] The first paragraph of arb-WL-1-154489-7725312#Arabic#system21#c.xml vs arb-WL-1-154489-7725312#Arabic#reference_1#r.xml.

Translation	Av. LT	Std. LT	Min. LT	Max. LT
Ref. 1	.873	.140	.590	1.447
Ref. 2	.851	.124	.636	1.256
Ref. 3	.838	.121	.657	1.177
Ref. 4	.865	.131	.639	1.429
Av. Ref.	.857	.124	.641	1.317
MT 1	.814	.110	.670	1.113
MT 2	.824	.109	.565	1.089
MT 3	.818	.113	.607	1.137
MT 4	.836	.116	.615	1.144
MT 5	.803	.097	.590	1.067
MT 6	.824	.116	.574	1.173
MT 7	.819	.115	.576	1.162
MT 8	.810	.104	.606	1.157
MT 9	.827	.114	.546	1.181
MT 10	.827	.122	.569	1.169
MT 11	.814	.116	.606	1.131
MT 12	.826	.112	.607	1.119
MT 13	.823	.115	.619	1.116
MT 14	.826	.115	.630	1.147
MT 15	.820	.107	.655	1.124
MT 16	.827	.112	.593	1.147
MT 17	.835	.117	.642	1.169
Av. MT	.822	.107	.623	1.106

Table 3: Average lexical tightness (Av. LT) for the reference vs machine translations, on the NIST Open MT 2008 Evaluation Arabic to English corpus. Standard deviation, minimum, and maximum values across the 120 texts are also shown.

in the MT-read version.

MT-raw vision came to me on dream in view of her dream: Arab state to travel to and group of friends on my mission and travel quickly I was with one of the girls seem close to the remaining more than I was happy and you're raised ended === known now

MT-read A vision came to me in a dream. I was to travel quickly to an Arab state with a group of friends on a mission. I was with one of the girls who seemed close to the remaining ones. I was happy and you are raised. It ended. It is known now.

Ref A Dream. My sister came to tell me about a dream she had while she slept. She was saying: I saw you preparing to travel to an Arab country, myself and a group of girlfriends. You were sent on a scholarship abroad, and you were preparing to travel quickly. You were with one of the girls, who appeared to be closer to you than the others, and I was happy and excited because you were traveling. The end. I now know !

The use of *vision* instead of *dream*, *state* instead of *country*, *friends* instead of *girlfriends*, *mission* instead of *scholarship*, *raised* instead of *excited*, along with the complete disapperance of *slept*, *sister*, *preparing*, *abroad*, all contribute to a dramatic loss of associative texture in the MT version. Highly associated pairs like *dream-slept*, *tell-saying*, *girlfriends-girls*, *travel-abroad*, *sister-girls*, *happy-excited*, *travel-traveling* are all missed in the machine translation, while the newly introduced word *raised* is quite unrelated to the rest of the vocabulary in the extract.

5 Experiment 3: Quality of Machine Translation

5.1 System-Level Comparison

In this experiment, we address the following question: Is it the case that when a worse MT system A and a better MT system B translate the same set of materials, B tends to provide more lexically tight translations?

To address this question, we use the Metrics-MATR 2008 development set (Przybocki et al., 2009) from NIST Open MT 2006 evaluation. Eight MT systems were used to translate 25 news articles from Arabic to English, and humans provided scores for translation adequacy on a 1-7 scale. We calculated the average lexical tightness over 25 texts for each of the eigth MT systems, as well as the average translation score for each of the systems. We note that human scores are available per text segments (roughly equivalent to a sentence, 249 segments in total for 25 texts), rather than for whole texts. We first derive a human score for the whole text for a given system by averaging the scores of the system's translations of the different segments of the text. We then derive a human score for an MT system by averaging the scores of its translations of the 25 texts. We found that the average adequacy score of a system is statistically significantly positively correlated with the average lexical tightness that the system's translations exhibit: r=0.630, n=8, df = 6, p<0.05.

5.2 Translation-Level Comparison

The same data could be used to answer the question: Is it the case that better translations are lexically tighter? Experiment 2 demonstrated that human reference translations are tighter than machine translations; does the same relationship hold for better vs worse machine translations? To address this question, 25 x 8 = 200 instances of (system, text) pairs can be used, where each has a human score for translation adequacy and a lexical tightness value. Human scores and lexical tightness of a translated text are significantly positively correlated, r=0.178, n=200, p<0.05. Note, however, that this analysis is counfounded by the variation in lexical tightness that exists between texts: As standard deviations and ranges in Tables 2 and 3 make clear, original human texts, as well as reference human translation for different texts, vary in their lexical tightness. Therefore, a lower lexical tightness value can be expected for certain texts even for adequate translations, while for other texts low values of lexical tightness signal a low quality translation. System-level analysis as presented in section 5.1 avoids this confounding, since all systems translated the same set of texts, therefore average tightness values per system are directly comparable.

6 Discussion and Conclusion

We presented a suggestive finding regarding the loss of associative texture in the process of machine translation, using comparisons between (a) original and back-translated texts, (b) reference and system translations, (c) better and worse machine translations. We represented the amount of association in a text using **word association profile** – a distribution of point wise mutual information between all pairs of content word types in a text. We used the average of the distribution, which we term **lexical tightness** – as a single measure of the amount of association in a text. We showed that the lexical tightness of human-composed texts is higher than that of the machine translated materials. While the phenomenon of the loss of associative texture has been theoretically predicted by translation scholars, lexical tightness is a computational measure capable of quantifying the extent of this phenomenon.

Our work complements that of Wong and Kit (2012) in demonstrating the potential utility of discourse-level phenomena to assess machine translations. First, we note that our findings are orthogonal to the main finding in Wong and Kit (2012) regarding loss of cohesion through insufficient word repetition, since our measure looks at pairs of word types, hence disregards repetitions. Second, the notion of pairwise word association generalizes the notion of lexical cohesive devices by looking not only at repeated reference with different lexical items or at words standing in certain semantic relations to each other, but at the whole of the lexical network of the text. Third, differently from the cohesion measure proposed by Wong and Kit (2012), the lexical tightness measure does not depend on lexicographic resources such as WordNet that do not exist in many languages.

References

Beata Beigman Klebanov and Michael Flor. 2013. Word Association Profiles and their Use for Automated Scoring of Essays. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August.

Antoine Berman. 2000. Translation and the Trials of the Foreign (translated from 1985 French original by L. Venuti). In Lawrence Venuti, editor, *The Translation Studies Reader*, pages 276–289. New York: Routledge.

Marine Carpuat and Michel Simard. 2012. The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada, June. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

David Graff and Christopher Cieri. 2003. English Gigaword LDC2003T05. Linguistic Data Consortium, Philadelphia.

Mark Przybocki, Kay Peterson, and Sebastien Bronsart. 2009. 2008 NIST metrics for machine translation (MetricsMATR08) development data.

Jörg Tiedemann. 2010. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden, July. Association for Computational Linguistics.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426, Montréal, Canada, June. Association for Computational Linguistics.

Lawrence Venuti. 1995. *The Translator's Invisibilty: A History of Translation*. London & New York: Routledge.

Billy Tak-Ming Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *EMNLP-CoNLL*, pages 1060–1068.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of the Machine Translation Summit XIII*.