The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean

Ray Clifford[1], Neil Granoien[1],

Douglas Jones[2], Wade Shen[2], Clifford Weinstein[2]

Defense Language Institute[1]
Monterey, California
{ray.clifford,neil.granoien}@us.army.mil

MIT Lincoln Laboratory[2]
Lexington, Massachusetts
{daj,swade,cjw}@ll.mit.edu

Abstract[•]

We report on initial experiments that examine the relationship between automated measures of machine translation performance (Doddington, 2003, and Papineni et al. 2001) and the Interagency Language Roundtable (ILR) scale of language proficiency/difficulty that has been in standard use for U.S. government language training and assessment for the past several decades (Child, Clifford and Lowe 1993). The main question we ask is how technology-oriented measures of MT performance relate to the ILR difficulty levels, where we understand that a linguist with ILR proficiency level N is expected to be able to understand a document rated at level N, but to have increasing difficulty with documents at higher levels. In this paper, we find that some key aspects of MT performance track with ILR difficulty levels, primarily for MT output whose quality is good enough to be readable by human readers.

Introduction

Current automated MT scoring techniques do not specifically consider the difficulty of input text in evaluating performance. We analyze the performance of MT with respect to input text difficulty and scoring methods. We focus our study on the behavior of the official NIST MT Evaluation scoring package based on the IBM BLEU scoring tool. We introduce a corpus of rated texts selected from five different languages with accompanying reference translations. Using the reference translations in this corpus, we conducted a variety of experiments that examine the difficulty-performance relationship. Some of the experiments address properties of the texts that may affect MT components (i.e., more difficult text may be more difficult to parse), whereas other experiments address MT performance in terms of NIST/BLEU scores.

"SPARK" Microcorpus Rated for ILR Difficulty

Language instructors for **S**panish, Farsi (**P**ersian), **A**rabic, **R**ussian, and **K**orean, at the U.S. Defense Language Institute (DLI) have selected and rated a small collection of documents at each of seven difficulty levels, across a range of topical domains, for the purpose of exploring the relationship between MT performance and input text difficulty. The Linguistic Data Consortium has agreed to make an online version of this corpus available to the MT research community. Each text is accompanied by at least four English reference translations and a commentary on the difficulty level.

Overview of the ILR Difficulty Levels

The ILR skill levels are an integral part of foreign language skill assessment in a variety of settings for agencies in the U.S. Government. A description of an ILR-based text classification scheme can be found in (Child et al., 1993 and Lowe 1999) and on the web (see References); some key points:

- **Level 1 texts**: contain short, discrete, simple sentences; generally pertain to the immediate time frame; often written in an orientational mode; require elementary level reading skill. **Example**: Newspaper announcements.
- **Level 2 texts**: convey facts with the purpose of exchanging information; do not editorialize on the facts; often written in an instructive mode; require limited working proficiency. **Example**: Newswire articles; TIDES/MT evaluation data.
- **Level 3 texts**: have denser syntax and highly analytic expressions; place greater conceptual demands on the reader; often written in an evaluative mode; may require the reader to 'read between the lines'; require general professional proficiency. **Example**: newspaper opinion / editorial articles.
- **Level 4 texts:** express creative thinking; assume a relative lack of shared personal information; often involve a highly individualized mode that projects the style of the author; require advanced professional proficiency. **Example**: essays; political editorials that reformulate social, economic or political policy.

Figure 1 shows a sampling of Spanish, Farsi, Arabic, Russian and Korean text segments in the SPARK corpus. To save space, only one example of each text difficulty level is shown for each of the seven levels in our corpus [1, 1+, 2, 2+, 3, 3+, 4]. Some basic statistics about the corpus are shown in Figure 2.

Arabic – Level 1 (Car Sale Advertisement)	
Src	قراي‌س عارش بـولـطم
Ref	A car needed for purchase
MT	required buying a car

• This work was sponsored by the Defense Language Institute under Air Force Contract number F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Korean – Level 1+ (Newspaper Article)	
Src	영국의 가디언지가 이 연구소의 보고서를 인용, **22**일 보도한 내용에 따르면 중국어 사용자는 **11**억 **2**천 **3**백만명에 달하는 것으로 나타났다.
Ref	According to the report, issued on the 22nd, the number of people who use Chinese has reached 1,123,000,000 (quoted from the British magazine, The Guardian).
MT	Great Britain D ¾ð price of land refers the report of the laboratory, 22 works in the contents which it reports tta_lu_myen the Chinese user in 1123000000 people tal_ha_nun with the thing ³ªÅ¸³µ´Ù.

Arabic – Level 2 (Newspaper Article)	
Src	برض يذلا فيعلا لازلزلا يلتق ددع عفترا وكيسكم قمصاعلا و ةيكيسكم تايالو عست احيرج 160 و اصخش 33 يلإ يتيس .
Ref	The number of dead people in the fierce earthquake that hit nine Mexican states and the capital Mexico City increased to 33 persons in addition to 160 wounded.
MT	rose number the violent earthquake which hit nine and prevented Mexican states fries Mexico City to 33 persons Ù ,H1 (B60 a casualties.

Spanish – Level 2+ (Newspaper Article)	
Src	La influencia europea se acentúa en el siglo XVII y XVIII y, sin embargo, entretejida a estas visiones surge la cultura africana con sus músicas y sus ritos.
Ref	The European influence becomes more obvious in the 17th and 18th Centuries, but nevertheless, intertwined in these visions emerges the African culture with its music and its rituals.
MT	The European influence is accentuated in century XVII and XVIII and, nevertheless, entretejida to these visions arise the African culture with its musics and their rites.

Russian – Level 3 (Political Commentary)	
Src	**У каждой нации есть своя великая национальная идея, она одна и та же - купить, обольстить или организовать весь мир, возглавить его и сделать так, чтобы все жили в соответствии с устоями, ценностями и приоритетами этой самой нации**
Ref	You see, each and every nation has it's own national idea, and it's always the same, to buy, to delude or to reorganize the whole universe, to seize the power and force all the people live according to the standards set by the nation.
MT	In each nation is its great national idea, it one and the same - to purchase, to flatter or to organize the entire world, to head it and to make so that all would live in accordance with the abutments, the values and the priorities of this nation itself.

Farsi – Level 3+ (Political Commentary)	
Src	یفسلف یاه هشیر بابردرظن لها هچرگ ، ددجت یسایسو یداصتقا،یبهذم، قافتا نآزاغآ یخیرات مییوقتءمرابردزینو هدس:دنقفتم لوق کیرد یگلمجام،دنرادنرظن یالطرص ع یدالیم مهدزونات مهدزناش یاه .دوب بیرغرد ددجت جاورودشر
Ref	Although there has not been consensus among scholars on philosophical, religious, economic, and political roots of modernization, they all collectively agree that the era between sixteenth to nineteenth centuries was the golden age of modernization.

Arabic – Level 4 (Political Commentary)	
Src	كمعظم أبناء جيلي، وثبت من القرية إلى أكبر عاصمة عربية بلا أية مقدمات، فكان من السهل علي أول من يصادفني أن يخدعني، تقدم باتجاهي على الرصيف رجل أيق، به شبه من نجوم السينما في الستينات، وهمس بأذني أن معه خاتماً ،همهاد دق ذهبياً يريد أن يبيعه بسعر زهيد، لأن ظرفاً مفاجئاً .روفلا یلع تلبقو
Ref	As the majority of my generation has proven, whether coming from a village or the largest city in the Arab World, and without any introduction, it was easy for the first person that saw me in the city to take advantage of me, as all of us experienced at one moment in our lifetime.
MT	as most my generation sons,, was who chances me of the on first plain deceives me, A elegant man presents with him on the quay in my direction quasi from the cinema stars in sixties, and a whisper in my ear with him a golden ring want cause to sell him in a petty price a circumstance Sudden had him, accepted immediately.

Figure 1: SPARK Corpus Sample

Reference Words					
	Spa	**Far**	**Ara**	**Rus**	**Kor**
L1	225	149	149	288	327
L1+	559	262	317	407	437
L2	1,001	508	643	832	632
L2+	1,317	957	1,285	1,276	1,051
L3	1,831	1,048	1,177	1,368	1,728
L3+	3,319	1,064	2,805	1,824	1,545
L4	2,208	1,483	2,920	1,339	2,215
Total	10,460	5,471	9,296	7,334	7,935
Source Words					
	Spa	**Far**	**Ara**	**Rus**	**Kor**
Total	10,098	4,806	6,536	6,110	4,839

Figure 2: SPARK Corpus Statistics

The text passage size generally increases with difficulty. Since there are four passages at each level, there are fewer words per level in the lower difficulty levels in the current version of the SPARK corpus.

Three Preliminary Experiments

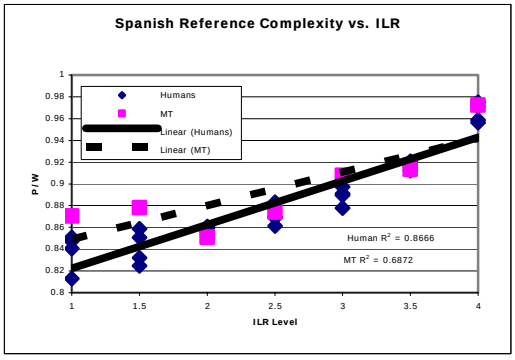
Our preliminary experiments address two different ways that input difficulty may affect MT performance. First, there may be direct effects of higher input text complexity that may have ripple effects on the MT system at the component level (language modeling, parsing, etc). Second, there may be unspecified system-internal effects which may be observable by comparing NIST/BLEU scores of the MT output with the input text difficulty.

The first experiment compares the complexity of parses of the reference texts against the source difficulty levels. The second experiment compares NIST/BLEU scores with ILR levels at the output level. The third experiment scores the human reference translations against themselves, the idea being that more divergent reference translations indicate a more complex translation space. Each of these experiments shows an interesting effect associated with ILR text difficulty.

1. ILR Difficulty and Reference Complexity

In the first experiment, we tagged and parsed[i] the English reference translations to probe the complexity of the source texts. Since the corpus is aligned at the sentence level, we felt that our assumption that the structural complexity of the reference translations is mirrored in the source sentences was acceptable (if not, it is certainly representative of the complexity of the targets an MT system is asked to produce).

For this experiment, we used the very simple measure of parse-nodes per word (P/W). P/W gives us a measure of the amount of structure that each word requires to generate it (i.e., its syntactic density, see (Jones and Rusk 2000) for similar linguistic probes). From a purely informational perspective, a higher P/W measure implies that for each target word, an MT system would need to generate more non-surface structural components to support it. The Spanish reference texts showed the strongest positive correlation of increasing P/W with increasing ILR difficulty. This result is shown in Figure 3, R^2=0.87 (unnormalized mean, across four translators, R^2=0.9683 for the best individual). When the MT is parsed and analyzed for P/W, there is also a positive, but weaker, relation: R^2=0.69 (with one MT system). Furthermore, the NIST scores for a machine translation of these sentences shows a negative correlation – more structurally complex reference translations are associated with a degradation in MT quality as measured by NIST scores, also shown in Figure 3.



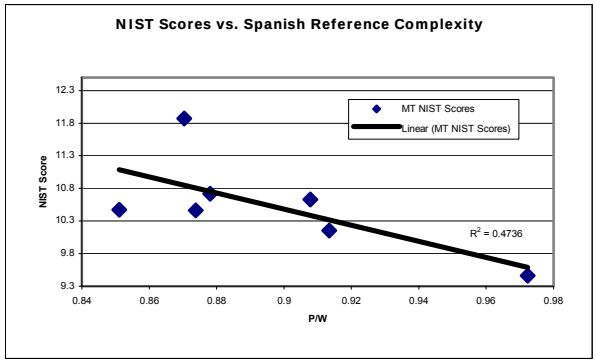


Figure 3: Structural Complexity

The reference complexity of the other languages was less related, R^2: Farsi=0.0495, Arabic = 0.5606, Russian=0.5313, Korean=0.381. The MT complexity was: Arabic=0.3104, Russian=0.2479, Korean=0. 6189.

2. ILR Difficulty and NIST MT Scores

Inspection of the MT output for these experiments reveals that Spanish texts were most intelligibly translated, whereas counterparts in Korean and Arabic was generally unintelligible (with Russian somewhere in between -- we did not have Farsi MT output for these experiments). For the purpose of situating the relative quality of our MT output across languages, we examine a boundary condition on intelligibility.

A Scrambling Baseline for MT Quality

Even perfectly translated words are not intelligible to the human reader if their order is permuted so that no ngrams > 1 have matches, i.e., "scrambled". So for each of the languages, we established a "scrambling baseline", and noticed that output that fell below this level had essentially hit the floor in terms of intelligibility.

In our second experiment, we compared NIST scores with ILR text difficulty. Only the Spanish MT output was above its scrambling baseline; the Arabic, Russian and Korean were below it, showing no clear association between ILR difficulty and NIST scores. The Spanish NIST scores are shown in Figure 4, where we see a general downward trend as difficulty increases.

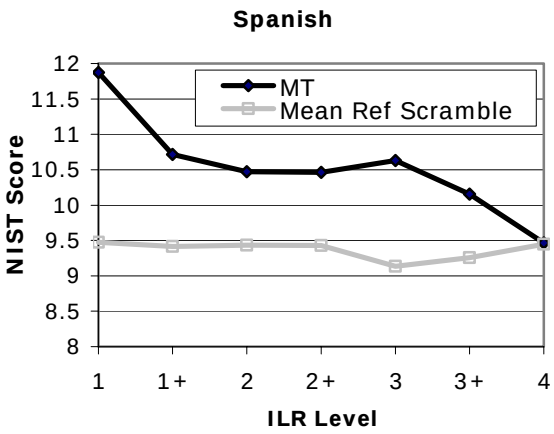


Figure 4: ILR Ratings and NIST Scores

3. ILR Difficulty and Reference Divergence

The third experiment was designed to probe the idea that more difficult texts are characterized by an increased level of conceptual content and implied world knowledge. In this sense, a reader of higher level text needs to apply more implicature and information than what is directly conveyed in the text itself to be able to understand its content. The effect of increased conceptual content and difficulty has unclear ramifications for machine translation. We might guess that, like humans, increased ILR difficulty may require more conceptual knowledge than that which is embodied in a transfer-based or statistical MT system. If true, we would expect lower levels of performance from an MT system with increasing ILR level. Taking another view, it may be the case that the translations of higher level texts do not require application of deep conceptual knowledge, that deep information can be preserved even through a shallow MT process.

Furthermore, it may be the case that the increase in conceptual density leads to a larger translation space, the range of possibilities for translations into the target language. If this is true, then we would expect that the range of human translations at higher levels would be wider. To get at this question further, we measure the similarity of held out reference translations against the remaining set in each of the SPARK languages as a function of ILR level. These results are shown in Figure 5. The general downward trends in NIST scores for reference translations scored against themselves indicates the increase in reference translation divergence we suspected might exist.

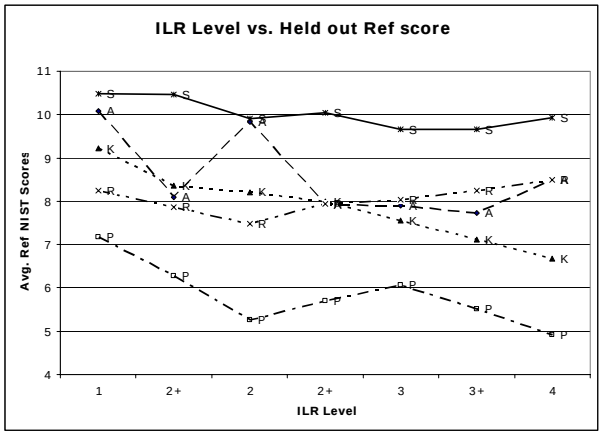


Figure 5: Held out reference scores

Conclusions and Future Work

Previous research on machine translation performance has not specifically taken into account the difficulty of the input text. We have shown that there are interesting and significant relationships between input text difficulty and various measures of machine translation performance. Two reasons why the effects are sometimes subtle and varied are (1) the SPARK corpus is relatively small, particularly at the lower ILR levels and (2) the quality of the MT output was relatively poor in these experiments. In our future work, we will explore our "scrambling baseline" against the well-studied data in the NIST MT-02 and MT-03 evaluations.

An important consideration, not only for these preliminary experiments but also for future work with larger corpora and better MT quality, is that what is difficult for people might not be difficult for machines and vice versa. Gaining a better understanding of these specific difficulties is one of our main goals for future work. To gain additional insight, we will continue the general inquiry of how the technology-centric measures compare with standard measures of human foreign language proficiency with additional measures of performance, such as the Defense Language Proficiency Test.[ii]

References

Doddington, G. (2003) "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics". NIST MT Web Site, February 2003. http://www.nist.gov/speech/tests/mt/

Brill, Eric (1995) "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics*. December, 1995.

Child, James R., Ray T. Clifford and Pardee Lowe, Jr. (1993) "Proficiency and Performance in Language Testing". Applied Language Learning, Vol 4.

Michael Collins (1997) "Three Generative, Lexicalised Models for Statistical Parsing". Proceedings of the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL), Madrid.

Defense Language Institute Course Catalog: ILR Skill Levels: http://www.monterey.army.mil/atfl/daa/skill.htm

Jones, Doug and Greg Rusk (2000) Toward a Scoring Function for Quality Driven Machine Translation. Proceedings COLING 2000.

Lowe, Pardee (1999) "James R. Child's Text Modes and Their Derivatives: A Compilation of Description", manuscript from course materials.

Papineni, Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu (2001) "Bleu: a Method for Automatic Evaluation of Machine Translation" IBM Computer Science Research Report RC22176 (W0109-022) 9/17/2001 http://domino.watson.ibm.com/library/

[i] For this experiment, we used Eric Brill's part of speech tagger (Brill 1995) and Michael Collins parser (Collins 1997).

[ii] We gratefully acknowledge helpful discussions with Dr. Martha Herzog, Ms. Sabine Atwell, Mr. James Dirgin, Dr. Jurgen Sottung and Mr. Michael Emonts at the Defense Language Institute in support of our research.