Ground Truth, Reference Truth & "Omniscient Truth" -- Parallel Phrases in Parallel Texts for MT Evaluation

M. Vanni* **C.R. Voss*** **C. Tate*** [§]

*Multilingual Computing Group
Army Research Lab
Adelphi, MD

[§]Dept. of Mathematics
University of Maryland
College Park, MD

{mvanni | voss }@arl.army.mil, ctate@math.umd.edu

Abstract

Recently introduced automated methods of evaluating machine translation (MT) systems require the construction of parallel corpora of source language (SL) texts with human reference translations in the target language (TL). We present a novel method of exploiting and augmenting these resources for task-based MT evaluation, assessing how accurately people can extract *Who*, *When*, and *Where* elements of information from TL output texts of different MT engines. This paper reports on the first phase of our research establishing a baseline MT evaluation process with (i) the construction and (ii) the annotation and inter-annotator rates of *an annotated extraction corpus*, and (iii) our results applying the corpus in the evaluation of three Arabic-to-English MT engines. In this corpus, the elements of interest are identified as parallel phrases across the parallel texts of the SL, the reference translations, and the MT engine outputs, where they are annotated and called, respectively the Ground Truth (GT), Reference Truth (RT), and Omniscient Truth (OT) items in the parallel texts. Our evaluation of three MT engines with the corpus yields precision and recall accuracy measures that, together with a loss measure, clearly rank the engines and, unlike other evaluation metrics, indicate diagnostically where output improvements will assist on extraction.

1 Introduction

Current methods of evaluating machine translation (MT) systems are costly: they require the construction of parallel corpora of source language (SL) texts with human reference translations in the target language (TL) prior to the run-time evaluations. We present a novel method of exploiting and augmenting these resources that we use for an experiment in task-based MT evaluation, assessing how accurately people can extract *Who*, *When*, and *Where* elements of information from TL output texts of different MT engines.

Our research approach is to divide into three stages, the analysis of which "end-to-end" MT engine-with-user combination produces the most complete and accurate information. First, we evaluate the MT output standalone (that will later be shown to users) for how adequately the engines preserve the content of the *Who*, *When*, and *Where* elements. Second, we conduct an experiment with users viewing the MT outputs of different engines and evaluate their responses (that they provide via our software tools) for how effectively they can extract the elements. Then, we use the results of these evaluations within a generalized linear model to test the relation of MT engine, document and subject variables in predicting the "end-to-end" MT engine-with-user accuracy in extracting the elements from MT output.

This paper reports on the first phase of the research approach with (i) the construction and (ii) the annotation, with inter-annotator rates, of ***an annotated extraction corpus***, and (iii) our results applying the corpus in the evaluation of three Arabic-to-English MT engines. In this corpus, the elements of interest are identified as parallel phrases across the parallel texts of the SL, the reference translations, and MT engines' outputs, where the elements are annotated and called, respectively the ground truth (GT), reference truth (RT), and omniscient truth (OT) items in the texts. Our evaluation of the three MT engines with the corpus yields precision and recall accuracy measures that, together with a loss measure, clearly rank the engines and, unlike other evaluation metrics, indicate diagnostically where output improvements will assist on extraction.

2 Approach

The construction of the annotated extraction corpus, illustrated in Figure 1, involves building the parallel texts, annotating them for the parallel phrases, and then augmenting the phrases in the MT output files with a higher-order, backoff categorization for evaluating the OT items in those files.

2.1 Parallel Texts

The corpus that we have created is effectively a three-way parallel corpus of the source language texts, reference translations, and MT outputs, aligned at the sentence level. We started with a collection of online Arabic language documents built by one native Arabic speaker with news article from ten different websites, where each article was selected for one of the who/when/where extraction tasks of the second stage of our research.

Four native Arabic speakers (including the one who built the collection), all bilingual in Arabic and English, then translated the documents into English to create the four reference translations for the corpus. We followed the guidelines established at the Linguistic Data Consortium for directing these individuals to create translations that preserve the full content of the documents as closely as possible and that do not add extra information which is not literally present in the text. They were instructed to translate the Arabic text on a sentence-by-sentence basis, creating English sentences that are fluent and do not contain Arabic constructions, such as sentences that start with the word "And" after the initial paragraph sentence.

To create the MT output files of the corpus, we ran the online Arabic documents through each of the three

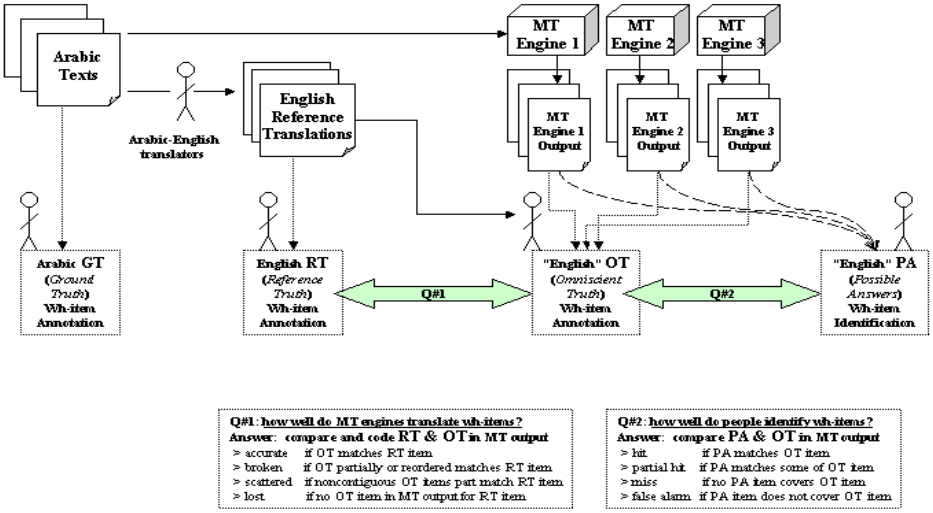

Figure 1. Process of Constructing an Annotated Extraction Corpus

Arabic-to-English MT engines that we had available in their most recent release as of the end of October 2003. As needed, we converted the documents into the input format required by the MT engine. We also opted to run the MT engines with any settings left at their default value.

2.2 Parallel Phrases

Given our second-stage goal of evaluating how well people can extract *Who, When,* and *Where* elements of information from MT output for the purpose of ranking the "end-to-end" MT engine-with-user combinations, we experimented with defining these elements at different levels of granularity. The key was to determine the most straightforward, non-technical description of "chunks" of information in "noisy" MT output[1] that the people in our experiments, who were neither translators nor linguists, would be able to detect readily without extensive training.

We started out examining the category descriptions for PER (person), ORG (organization), LOC (location), and TIMEX (time expression) in the ACE program guidelines. Reading these guidelines and effectively learning the large and fine-grained distinctions among the categories that are extensively documented with examples requires several hours. Furthermore the categories are defined over the smallest atomic element of information, not the phrasal or chunk level that we needed in order to assess both the content of the MT output in the first phase of our work, and the feasibility of people extracting *Who, When,* and *Where* elements from the noisy MT output in the second phase of our work.

As a result, we established instead intuitive semantic descriptions, where the chunk could include attributes if that information was local within the syntactic phrase in the SL or reference translations. We pre-tested and refined the descriptions on members of our staff with no linguistic training, after giving them about twenty minutes training.

The identification of *Who, When,* and *Where* ground truth (GT) elements in the Arabic texts was set by one native Arabic translator and then vetted by a trained linguist in possession of the English human reference translations, who then marked up these documents for their parallel reference translation (RT) phrases.

The "who" category of our annotations consists of mentions of individual persons or groups of people, organizations, corporations, governments or other entities functioning as persons in the context of the SL passage. Here we include roles, names, objects with human identity and numbers referring to persons. The "where" category is comprised of names, proper/common nouns, and expressions such as prepositional phrases which refer to locations, regions, facilities, civil structures and other bounded geographic areas. The "when" category contains time and date expressions with standard proper noun month-day-year references, common nouns referring to time periods or instants, unique identifiers for temporally-defined events, or prepositional phrases referring to specific time periods.

After the GT-RT annotations were established, we developed the following procedure for identifying the corresponding "omniscient truth" (OT) elements in the MT outputs. Given a listing of the RT elements by document in order of appearance within each sentence of the document, the annotators searched within the same sentence of the MT output text for the OT that best approximated the RT element. The OT "chunks" were selected semantically by the annotators, so that even when they found incorrect English syntax or incomplete translations only roughly corresponding to the RT element, they could identify an OT item. The set of OTs for a document vary with the MT engine that generated the output text in the document. This can be seen in the example in Figure 2 where the underlined subject of the verb is translated by MT3, is transliterated by MT1 and MT2, and is separated across the verb in MT1.

[1] "Noisy" MT output refers to text output by MT engines that contains ungrammatical phrases with words out of order, incorrect or peculiar word selections, unrecognizable transliterated names, SL words left untranslated, and so on.

GT: كتبت ريم الميح: في قصر بيان....
RT: Reem Meeh wrote:
yesterday at Bayan Palace...
MT1: Reem wrote 'Lmeeaa : [S]
in a statement derelict , ...
MT2: I wrote Rim almyai [المبح] : [A]
in the short statement, ...
MT3: clerks move flowing : [Z]
in castle demonstration/statement? ...

Figure 2. Sample Parallel Phrases in Parallel Texts: *Who* ground truth (GT) phrase in Arabic source text, *Who* reference truth (RT) phrase in reference translation, *Who* omnisicient truth (OT) phrases with backoff codes in output texts of three Arabic-to-English MT engines

2.3 MT Output Backoff Classification

As annotators were reading the MT output texts and identifying the OT items to be marked up, they spontaneously started categorizing the patterns of errors in the MT output that directly affected their decision-making process of establishing the boundaries of an OT item. For example, in Figure 2, they designated the open class words such as "wrote" in the MT1 text that appear incorrectly inside of the translated phrases as "trapped words." As the markup process continued, the name for the OT items with such trapped words inside evolved into "split items."

When we observed that the annotators were regularly using their terms for error patterns to resolve differences in their OT markups, we realized this information was central to the OT identification process and decided to codified it by grouping their error analysis patterns into four classification categories (A, B, S, and Z) and then tested their consistency in assigning the classification labels to the OT items.

Definitions of OT Item Classifications

A: 1) Exact match, synonym, or paraphrase
2) Contiguous phrase
3) Words in grammatical word order

B: 1') Exact match, synonym, paraphrase
OR partial match with some content loss
2) Contiguous phrase
3') Words in grammatical word order
OR out of grammatical order

S: 1') Exact match, synonym, paraphrase
OR partial match with some content loss
2') Non-contiguous phrase
3') Words in grammatical word order
OR reordered OR out-of-order

Z: Lost OR not recognizable

The OT identification and backoff classification process worked as follows. First, the annotators would compare their respective OT items with the RT items for a match, within the relevant sentence, that preserved the RT meaning and that formed a grammatical element. These items were the best, or "A" cases. When there was no evidence for that form of an OT item in the MT output, they would do a backoff analysis and look for a chunk of contiguous words that *would* be a good OT item, *if* the words were re-arranged or had another word or two added in. Since these items were clearly not as easy to detect because they required spotting the relevant words in a partial or noisy pattern, these items became "B" cases. The split items, mentioned earlier, became the "S" cases. Finally, for those cases where the annotators could not identify any text in the relevant MT output sentence that conveyed the name in or the semantic content of the RT item (as occurs in the MT3 output for the subject's name "Reem Meeh" in Figure 2), the annotators designated that item "Z" to record that it was lost in translation.

3 Results

3.1 Backoff Classification

We evaluated the inter-annotator agreement rates on the ABSZ coding with the Kappa statistic (Cohen, 1960) for each MT engine, both across and within the who/when/where types after one round of annotation, but before the final resolution of the codes. The scores were all within the 0.6 to 0.8 "good agreement" range. Also, three of the nine Kappa scores for within who/where/when types were above 0.8 in the "very good" range. Most of the differences among annotators were at the A-B boundary.

The results of assigning each OT item to one of the four categories, A, B, S, or Z, are shown in Table 1. The total rows for each of the MT engines indicate, across who/when/where elements, how many are categorized as OTs (As, Bs, or Ss) and how many are lost in translation (Zs). The precision measure is the number of As divided by the number of OTs, and the recall measure is the number of As divided by the number of RTs. The loss measure is the number of Zs divided by the number of RTs. RT totals used in the Recall calculations for all MT engines are: 156 for all wh-items, 56 for Who items, 56 for Where items, and 44 When items.

	Backoff Classification					**Accuracy Measures**		
	A	B	S	Z	OT	**Prec**	**Rec**	**Loss**
MT1Total	67	51	**20**	**18**	**137**	**.49**	**.43**	**.12**
Who	21	17	12	6	50	.42	.38	.11
Where	34	15	2	5	51	.67	.61	.09
When	12	19	6	7	36	.33	.27	.16
MT2Total	**91**	49	9	7	**149**	**.61**	**.58**	**.05**
Who	29	19	7	1	55	.53	.52	.02
Where	41	12	1	2	54	.76	.73	.04
When	21	18	1	4	40	.53	.48	.09
MT3Total	67	**75**	4	10	**146**	**.46**	**.43**	**.06**
Who	21	26	2	7	49	.43	.38	.13
Where	33	22	0	1	55	.60	.59	.02
When	13	27	2	2	42	.31	.30	.05

Table 1. Counts of ABSZ Codes and Precision/Recall/Loss Percentages on MT Output of Annotation Extraction Corpus

3. 2 Interpretation

The precision, recall, and loss measures in Table 1 serve to tease apart the differences among the three Arabic-English MT systems that we tested. There are four results in this table. First, notice the substantially higher precision and recall scores of MT2 (.61 and .58), compared to those of MT1 (.49 and .43) and MT3 (.46 and .43), based on "A" scores. Second, while the precision and recall scores for MT1 and MT3 nearly identical, the loss scores based on "Z"s make is clear that MT1 is much weaker in preserving content. Third, MT1 is also weaker in preserving phrasal integrity, with more than twice the number of "S" split phrases in the output compared to the other two engines. Finally, Table 1 also makes clear that MT3 is the mostly likely engine to output less-than-correct "B" partial or broken-syntax translations.

To recap, the results of our work so far indicate we can rank the MT engines in our study on their accuracy and throughput in translating the wh-elements of interest for later extraction: MT2 provides the strongest overall results, MT1 has the weakest overall results because of its loss of content and phrasal integrity, and MT3 falls between the other two, with accuracy below that of MT2 but with better content throughput than MT1.

4 Related Work

We have developed a novel two-part approach to standalone MT engine evaluation that augments parallel text resources into an annotated extraction corpus and applies it in a focused *Who*, *When*, and *Where* backup classification of MT output text. This two-part approach is comparable to other current annotate-and-train/test approaches found in the processing of natural language texts for a wide range of applications, such as (i) tagged corpora for information extraction (Sundheim, 1991), (ii) bracketed corpora for parsing (Marcus, *et al*., 1993), and (iii) sense-tagged corpora for word sense disambiguation (Kilgariff and Palmer, 1999), to name but a few. These applications first require constructing corpora, developing well-documented annotation procedures for human annotators, determining the inter-annotator agreement rates, and resolving final annotations on the corpus. For many NLP applications, the annotated corpora then serve to train/test the algorithm for automating a particular task. In our work reported here, the annotated extraction corpus has served to develop the backoff classification algorithm for MT evaluation.[2]

While others have made *unannotated* parallel bilingual corpora central to their MT evaluation research[3], it is not yet clear what the results from these automated metrics signify. For example, Hovy and Ravichandran (2003) have shown that MT output that outperforms reference translations on these metrics may nevertheless be incomprehensible to human readers. Our approach with parallel corpora *annotated* for *Who*, *When*, *Where* extraction will allow us to test, in the second stage of our research, for a predictive model that can cross-validate our backoff evaluation performance measures with the effectiveness measures achieved by MT engine-with-user combinations carrying out extraction tasks.

[2] The corpus is also used in the second stage of our research on task-based extraction evaluation, not detailed in this paper.

[3] Papineni, *et al*. (2002), Doddington (2002).

5 Conclusions and On-Going Work

This paper reports on the first phase of our research establishing a standalone MT evaluation process with (i) the construction and (ii) the annotation and inter-annotator rates of *an annotated extraction corpus*, and (iii) our results applying the corpus in the evaluation of three Arabic-to-English MT engines. In this corpus, the elements of interest are identified as parallel phrases across the parallel texts of the SL, the reference translations, and the MT engine outputs, where they are annotated and called, respectively the Ground Truth (GT), Reference Truth (RT), and Omniscient Truth (OT) items in the parallel texts. Our evaluation of three MT engines with the corpus yields precision and recall accuracy measures that, together with a loss measure, clearly rank the engines and, unlike other evaluation metrics, indicate diagnostically where output improvements will assist on extraction.

We are currently conducting analyses, as part of the second stage of this research, on the results of task-based categorization, extraction, and template-completion experiments, where people read output text from the same three MT engines reported on in this paper. Given the results from the backoff classification found so far, we hypothesize that people will work most effectively with MT2 output. We also predict that there will be a range of individual differences in how well people are able to carry out these tasks on the output of MT1 and MT 3, as a function of how much experience they have with MT output.

Acknowledgements. This work was supported by the Center for Advanced Study of Language (CASL) at the University of Maryland, College Park, and the US DOD. We acknowledge the support of our lead translator, Jamal Laoudi, who supervised the translation and ground truthing of the document collection, and two other team members, Sooyon Lee and Joi Turner, who contributed to the OT and ABSZ annotation development and resolution steps.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*: 20 (pp. 37-46).

Doddington, G. (2002) "Automatic evaluation of machine translation quality using n-gram co-occurence statistics." In *Proceedings of HLT 2002*, Human Language Technology Conference, San Diego, CA.

Hovy, E. and D. Ravichandran (2003). Holy and Unholy Grails. Presentation at Panel, "Have we found the holy grail?" MT Summit IX, New Orleans, LA.

Kilgariff, A. and M. Palmer (1999). *Computers and the Humanities*: 34:1-2 (Special issue on Senseval1).

Marcus, M. B. Santorini, and M. Marcinkiewicz (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*: 19.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, PA

Sundheim, Beth, ed. (1991). In *Proceedings of the Third Message Understanding Conference* (MUC-3), San Diego, California. Morgan Kaufmann, San Mateo, CA.