The Feasibility of HMEANT as a Human MT Evaluation Metric

Alexandra Birch a.birch@ed.ac.uk
Barry Haddow bhaddow@inf.ed.ac.uk
Ulrich Germann ugermann@inf.ed.ac.uk
Maria Nadejde maria.nadejde@gmail.com
Christian Buck cbuck@lantis.de
Philipp Koehn pkoehn@inf.ed.ac.uk

University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB, UK

Abstract

There has been a recent surge of interest in semantic machine translation, which standard automatic metrics struggle to evaluate. A family of measures called MEANT has been proposed which uses semantic role labels (SRL) to overcome this problem. The human variant, HMEANT, has largely been evaluated using correlation with human contrastive evaluations, the standard human evaluation metric for the WMT shared tasks. In this paper we claim that for a human metric to be useful, it needs to be evaluated on intrinsic properties. It needs to be reliable; it needs to work across different language pairs; and it needs to be lightweight. Most importantly, however, a human metric must be discerning. We conclude that HMEANT is a step in the right direction, but has some serious flaws. The reliance on verbs as heads of frames, and the assumption that annotators need minimal guidelines are particularly problematic.

1 Introduction

Human evaluation is essential in machine translation (MT) research because it is the ultimate way to judge system quality. Furthermore, human evaluation is used to evaluate automatic metrics which are necessary for tuning system parameters. Unfortunately, there is no clear consensus on which evaluation strategy is best. Humans have been asked to judge if translations are correct, to grade them and to rank them. But it is often very difficult to decide how good a translation is, when there are so many possible ways of translating a sentence. Another problem is that different types of evaluation might be useful for different purposes. If the MT is going to be the basis of a human translator's work-flow, then post-editing effort seems like a natural fit. However, for people using MT for gisting, what we really want is some measure of how much meaning has been retained.

We clearly need a metric which tries to answer the question, how much of the meaning does the translation capture. In this paper, we explore the use of human evaluation metrics which attempt to capture the extent of this meaning retention. In particular, we consider HMEANT (Lo and Wu, 2011a), a metric that uses semantic role labels to measure how much of the "who, why, when, where" has been preserved. For HMEANT evaluation, annotators are instructed to identify verbs as heads of semantic frames. Then they attach role fillers to the heads and finally they align heads and role fillers in the candidate translation with those in a reference translation. In a series of papers, Lo and Wu (2010, 2011b,a, 2012) explored a number of questions, evaluating HMEANT by using correlation statistics to compare it to judgements of human adequacy and contrastive evaluations. Given the drawbacks of those evaluation measures, which we discuss in Sec. 2, they could just as well have been evaluating the human adequacy and contrastive judgements using HMEANT. Human evaluation metrics need to be judged on other intrinsic qualities, which we describe below.

The aim of this paper is to evaluate the effectiveness of HMEANT, with the goal of using it to judge the relative merits of different MT systems, for example in the shared task of the Workshop on Machine Translation.

In order to be useful, an MT evaluation metric must be *reliable*, be *language independent*, have *discriminatory power*, and be *efficient*. We address each of these criteria as follows:

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 52–61,
Sofia, Bulgaria, August 8-9, 2013

Reliability We produce extensive IAA (Inter-annotator agreement) for HMEANT, breaking it down into the different stages of annotation. Our experimental results show that whilst the IAA for HMEANT is acceptable at the individual stages of the annotation, the compounding effect of disagreement at each stage of the pipeline greatly reduces the effective overall IAA — to 0.44 on role alignment for German, and, only slightly better, 0.59 for English. This raises doubts about the reliability of HMEANT in its current form.

Discriminatory Power We consider output of three types of MT system (Phrase-based, Syntax-based and Rule-based) to attempt to gain insight into the different types of semantic information preserved by the different systems. The Syntax-based system seems to have a slight edge overall, but since IAA is so low, this result has to be taken with a grain of salt.

Language Independence We apply HMEANT to both English and German translation outputs, showing that the guidelines can be adapted to the new language.

Efficiency Whilst HMEANT evaluation will never be as fast as, for example, the contrastive judgements used for the WMT shared task, it is still reasonably efficient considering the fine-grained nature of the evaluation. On average, annotators evaluated about 10 sentences per hour.

2 Related Work

Even though the idea that machine translation requires a semantic representation of the translated content is as old as the idea of computer-based translation itself (Weaver, 1955), it has not been until recently that people have begun to combine statistical models with semantic representations. Jones et al. (2012), for example, represent meaning as directed acyclic graphs and map these to PropBank (Palmer et al., 2005) style dependencies. To evaluate such approaches properly, we need evaluation metrics that capture the accuracy of the translation.

Current automatic metrics of machine translation, such as BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009) and TER (Snover et al., 2009b), which have greatly accelerated progress in MT research, rely on shallow surface properties of the translations, and only indirectly capture whether or not the translation preserves the meaning. This has meant that potentially more sophisticated translation models are pitted against the flatter phrase-based models, based on metrics which cannot reflect their strengths. Callison-Burch et al. (2011) provide evidence that automatic metrics are inconsistent with human judgements when comparing rule-based against statistical machine translation systems.

Automatic evaluation metrics are evaluated and calibrated based on their correlation with human judgements. However, after more than 60 years of research into machine translation, there is still no consensus on how to evaluate machine translation based on human judgements. (Hutchins and Somers, 1992; Przybocki et al., 2009).

One obvious approach is to ask annotators to rate translation candidates on a numerical scale. Under the DARPA TIDES program, the Linguistic Data Consortium (2002) developed an evaluation scheme that relies on two five-point scales representing fluency and adequacy. This was also the human evaluation scheme used in the annual MT competitions sponsored by NIST (2005).

In an analysis of human evaluation results for the WMT '07 workshop, however, Callison-Burch et al. (2007) found high correlation between fluency and adequacy scores assigned by individual annotators, suggesting that human annotators are not able to separate these two evaluation dimensions easily. Furthermore these absolute scores show low inter-annotator agreement. Instead of giving absolute quality assessments, annotators appeared to be using their ratings to rank translation candidates according to their overall preference for one over the other.

In line with these findings, Callison-Burch et al. (2007) proposed to let annotators rank translation candidates directly, without asking them to assign an absolute quality assessment to each candidate. This type of human evaluation has been performed in the last six Workshops on Statistical Machine Translation.

Although it is useful to have a score or a rank for a particular sentence, especially for evaluating automatic metrics, these ratings are necessarily a simplification of the real differences between translations. Translations can contain a large number of different types of errors of varying severity. Even if we put aside difficulties with selecting one preferred sentence, ranking judgements are difficult to generalise. Humans are shown five translations at a time, and there is a high cognitive cost to ranking these at once. Furthermore, these repre-

sent a subset of the competing systems, and these rankings must be combined with other annotators judgements on five other system outputs to compute an overall ranking. The methodology for interpreting the contrastive evaluations has been the subject of much recent debate in the community (Bojar et al., 2011; Lopez, 2012).

There has been some effort to overcome these problems. HTER (Snover et al., 2009a) is a metric which counts the number of edits needed by a human to convert the machine translation so as to convey the same meaning as the reference. This type of evaluation is of some use when one is using MT to aid human translation (although the relationship between number of edits and actual effort is not straightforward (Koponen, 2012)), but it is not so helpful when one's task is gisting. The number of edits need not correlate with the severity of the semantic differences between the two sentences. The loss of a negative, for instance, is only one edit away from the original, but the semantics change completely.

Alternatively, HyTER (Dreyer and Marcu, 2012) is an annotation tool which allows a user to create an exponential number of correct translations for a given sentence. These references are then efficiently exploited to compare with machine translation output. The authors argue that the current metrics fail simply because they have access to sets of reference translations which are simply too small. However, the fact is that even if one does have access to large numbers of translations, it is very difficult to determine whether the reference correctly captures the essential semantic content of the references.

The idea of using semantic role labels to evaluate machine translation is not new. Giménez and Màrquez (2007) proposed using automatically assigned semantic role labels as a feature in a combined MT metric. The main difference between this application of semantic roles and MEANT is that arguments for specific verbs are taken into account, instead of just applying the subset agent, patient and benefactor. This idea would probably help human annotators to handle sentences with passives, copulas and other constructions which do not easily match the most basic arguments. On the other hand, verb specific arguments are language dependent.

Bojar and Wu (2012), applying HMEANT to English-to-Czech MT output, identified a number of problems with HMEANT, and suggested a variety of improvements. In some respects, this work is very similar, except that our goal is to evaluate HMEANT along a range of intrinsic properties, to determine how useful the metric really is to evaluation campaigns such as the workshop on machine translation.

3 Evaluation with HMEANT

3.1 Annotation Procedure

The goal of the HMEANT metric is to capture essential semantic content, but still be simple and fast. There are two stages to the annotation, the first of which is semantic role labelling (SRL). Here the annotator is directed to select the actions, or frame heads, by marking all the verbs in the sentence except for auxilliaries and modals. The roles (or slot fillers) within the frame are then marked and each is linked with a unique action. Each role is given a type from an inventory of 11 (Table 1), and an action with its collection of corresponding roles is known as a *frame*. In the role annotation the idea is to get the annotator to recognise *who did what* to *who, when, where* and *why* in both the references and the MT outputs.

who	**what**	**whom**	**when**	**where**
agent	patient	benefactive	temporal	locative
why	**how**			
purpose	degree, manner, modal, negation, other			

Table 1: Semantic roles

The second stage in the annotation is alignment, where the annotators match elements of the SRL annotation in the reference with that in the MT output. The annotators link both actions and roles, and these alignments can be matched as "Correct" or "Partial" matches, depending on how well the action or role is translated. The guidelines for the annotators are deliberately minimalistic, with the argument being that non-experts can get started quickly. Lo and Wu (2011a) claim that unskilled annotators can be trained within 15 minutes.

In all such human evaluation, there is a trade-off between simplicity and accuracy. Clearly when evaluating bad machine translation output, we do not want to label too much. However, sometimes having so little choice of semantic roles can lead to confusion and slow down the annotator when more complicated examples do not fit the scheme. Therefore, common exceptions need to be handled either in the roles provided, or in the annotator guidelines.

3.2 Calculation of Score

The overall HMEANT score for MT evaluation is computed as the f-score from the counts of matches of frames and their role fillers between the reference and the MT output. Unmatched frames are excluded from the calculation together with all their corresponding roles.

In recognition that preservation of some types of semantic relations may be more important than others for a human to understand a sentence, one may want to weight them differently in the computation of the HMEANT score. Lo and Wu (2012) train weights for each role filler type to optimise correlation with human adequacy judgements. As an unsupervised alternative, they suggest weighting roles according to their frequency as approximation to their importance.

Since the main focus of the current paper is the annotation of the actions, roles and alignments that HMEANT depends on, we do not explore such different weight-setting schemes, but set the weights uniformly, with the exception of a partial alignment, which is given a weight of 0.5. HMEANT is thus defined as follows:

$$
\begin{aligned}
F_i &= \text{\# correct or partially correct fillers for PRED i in MT} \\
MT_i &= \text{total \# fillers for PRED i in MT} \\
REF_i &= \text{total \# fillers for PRED i in REF} \\
P &= \sum_{matched\ i} \frac{F_i}{MT_i} \\
R &= \sum_{matched\ i} \frac{F_i}{REF_i} \\
P_{total} &= \frac{P_{correct} + 0.5P_{partial}}{\text{total \# predicates in MT}} \\
R_{total} &= \frac{P_{correct} + 0.5P_{partial}}{\text{total \# predicates in REF}} \\
\text{HMEANT} &= \frac{2 * P_{total} * R_{total}}{P_{total} + R_{total}}
\end{aligned}
$$

3.3 Automating HMEANT

One of the main directions taken by the authors of HMEANT is in creating a fully automated version of the metric (MEANT) in (Lo et al., 2012). The metric combines shallow semantic parsing with a simple maximum weighted bipartite matching algorithm for aligning semantic frames. They use approximate matching schemes (Cosine and Jaccard similarity) for matching roles, with the latter producing better alignments (Tumuluru et al., 2012). They demonstrate that MEANT correlates with human adequacy judgements better than other commonly used automatic metrics. In this paper we focus on human evaluation, as it is essential for building better automatic metrics, and therefore a more fundamental problem.

4 Experimental Setup

4.1 Systems and Data Sets

We performed HMEANT evaluation on three systems selected from 2013 WMT evaluation[1]. The systems we selected were `uedin-wmt13`, `uedin-syntax` and `rbmt-3`, which were chosen to provide us with a high performing phrase-based system, a high performing syntax-based system and the top performing rule-based system, respectively. The cased BLEU scores of the three systems are shown in Table 2.

System	Type	de-en	en-de
`uedin-wmt13`	Phrase	26.6	20.1
`uedin-syntax`	Syntax	26.3	19.4
`rbmt-3`	Rule	18.8	16.5

Table 2: Cased BLEU on the full `newstest2013` test set for the systems used in this study

We randomly selected sentences from the en-de and de-en `newstest2013` tasks, and extracted the corresponding references and system outputs for these sentences. For the en-de task, 75% of our selected sentences were selected from the section of `newstest2013` that was originally in German, with the other 25% from the section that was originally in English. The sentence selection for the de-en task was performed in a similar manner. For presentation to the annotators, the sentences were split into segments of 12. We found that with practice, annotators could complete one of these segments in around 100-120 minutes. In total, with close to 70 hours of annotator effort, we evaluated 142 sentences of German, and 72 sentences of English. The annotation for each sentence includes 1 reference, 3 system outputs, and their corresponding alignments. Apart from 5 singly-annotated German sentences, and 1 singly-annotated English sentence, all sentences were annotated by exactly 2 annotators.

[1] `www.statmt.org/wmt13`

4.2 Annotation

The annotation for English was performed by 3 different annotators (E1, E2 and E3), and the German annotation by 2 annotators (D1 and D2). All the English annotators were machine translation researchers, with E1 and E2 both native English speakers whereas E3 is not a native speaker, but lives and works in an English-speaking country. The two German annotators were both native speakers of German, with no background in computational linguistics, although D2 is a teacher of German as a second language and has had linguistic training.

The HMEANT evaluation task was carried out following the framework described in Lo and Wu (2011a) and Bojar and Wu (2012). For each sentence in the evaluation set, the annotators were first asked to mark the semantic frames and roles (i.e., slot fillers within the frame) in a human reference translation of the respective sentence. They were then presented with the output of several machine translation systems for the same source sentence, one system at a time, with the reference translation and its annotations visible in the left half of the screen (cf. Fig. 1). For each system, the annotators were asked to annotate semantic frames and slot fillers in the translation first, and then align them with frame heads and slot fillers in the human reference translation. Annotations and alignment were performed with Edi-HMEANT[2], a web-based annotation tool for HMEANT that we developed on the basis of Yawat (Germann, 2008). The tool allows the alignment of slots from different semantic frames, and the alignment of slots of different types; however, such alignments are not considered in the computation of the final HMEANT score.

The annotation guidelines were essentially those used in Bojar and Wu (2012), with some additional English examples, and a complete set of German examples. For ease of comparison with prior work, we used the same set of semantic role labels as Bojar and Wu (2012), shown in Table 1. Given the restriction that the head of a frame can consist of only one word, a convention was made that all other verbs attached to the main verb such as modals, auxiliaries or separable particles for German verbs, would be labelled as *modal*. This was the only change we made to the HMEANT scheme.

[2]Edi-HMEANT is part of the *Edinburgh Multi-text Annotation and Alignment Tool Suite* (http://www.statmt.org/edimtaats).

5 Results and Discussion

5.1 Inter-Annotator Agreement

We first measured IAA on role identification, as in Lo and Wu (2011a), except that we use exact match on word spans as opposed to the approximate match employed in that reference. Whilst exact match is a harsher measure, penalising disagreements related to punctuation and articles, using any sort of approximate match would mean having to deal with N:M matches. IAA is defined as follows:

$$IAA = \frac{2 * P * R}{P + R}$$

Where P is defined as the number of labels (either heads, roles, or alignments) that match between annotators, divided by the total number of labels given by annotator 1. And R is defined the same way for annotator 2. This is similar to an F-measure (f1), where we consider one of the annotators as the gold standard. The IAA for role identification is shown in Table 3.

	Reference		Hypothesis	
Lang.	matches	f1	matches	f1
de	865	0.846	2091	0.737
en	461	0.759	1199	0.749

Table 3: IAA for role identification. This is calculated by considering exact endpoint matches on all spans (predicates and arguments).

The agreements in Table 3 are not too different from those reported in earlier work. We note that the IAA for the German annotators drops for the MT system outputs, but this may be because the English annotators (as MT researchers) are less bothered by bad MT output than their counterparts working on the German texts.

Next we looked at the IAA on role classification, the other IAA figure provided by Lo and Wu (2011a). We only considered roles where both annotators had marked the same span in the same frame, with the frame being identified by its action. The IAA for role classification is shown in Table 4.

Again, we show similar levels of IAA to those reported in (Lo and Wu, 2011a). Examining the disagreements in more detail, we produced counts of the most common role type disagreements, by

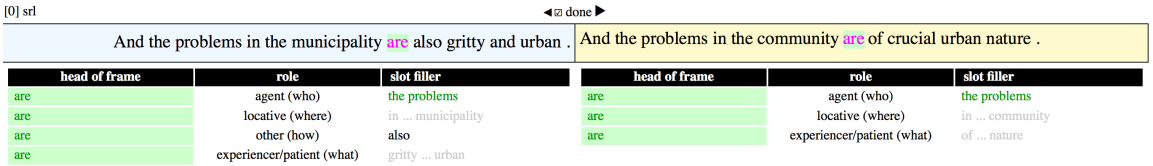

Figure 1: Example of a sentence pair annotated with Edi-HMEANT. The reference translation is on the left, the machine translation output on the right. Head and slot fillers for each semantic frame are marked by selecting spans in the text and automatically listed in tables below the respective sentences. Frames and slot fillers are aligned by clicking on table cells. The alignments of the semantic frames are highlighted: green (grey in black and white version) for *exact match* and grey (light grey) for *partial match*.

	Reference		Hypothesis	
Lang.	matches	f1	matches	f1
de	425	0.717	1050	0.769
en	245	0.825	634	0.826

Table 4: IAA for role classification. We only consider cases where annotators had marked the same span in the same frame.

Role 1	Role 2	Count
Agent	Experiencer-Patient	110
Degree-Extent	Modal	92
Beneficiary	Experiencer-Patient	45
Experiencer-Patient	Manner	26
Manner	Other	25

Table 5: Most common role type disagreements, for German

language. We show the top 5 disagreements in Tables 5 and 6. Essentially these show that the most common role types provide the most confusions.

In order to shed more light on the role type disagreements, we examined a random sample of 10 of the English annotations where the annotators had disagreed about "Agent" versus "Experiencer-Patient". In 7 of these cases, there was a definite correct answer, according to the annotation guidelines. Of the other 3, there were 2 cases of poor MT output making the semantic interpretation difficult, and one case of existential "there". Of the 7 cases where one annotator appears in error, 3 were passive, 1 was a copula, and 1 involved the verb

Role 1	Role 2	Count
Agent	Experiencer-Patient	44
Manner	Other	22
Degree-Extent	Temporal	12
Degree-Extent	Other	12
Beneficiary	Experiencer-Patient	11

Table 6: Most common role type disagreements, for English

"receive". For the other 2 there was no clear reason for the error. From this small sample, we suggest that passive constructions are still difficult to annotate semantically.

The last of elements of the semantic frames to be considered for IAA are the actions, i.e. the frame heads or predicates. In this case identifying a match was straightforward as actions are identified by a single token. The IAA for action identification is shown in Table 7.

	Reference		Hypothesis	
Lang.	matches	f1	matches	f1
de	238	0.937	592	0.826
en	126	0.818	362	0.868

Table 7: IAA for action identification.

We see fairly high IAA for actions, which seems encouraging, but given the importance of actions in HMEANT, we probably need the scores to be higher. Most of the problems with the identification of actions centre around multiple-verb constructions and participles.

We now turn our attention to the second stage of the annotation process where the annotators marked alignments between slots and roles. These provide the relevant statistics for the calculation of the HMEANT score so it is important that they are annotated reliably.

Firstly, we consider the alignment of actions. In this case, we use pipelined statistics, in that if one annotator marks actions in the reference and hypothesis, then aligns them, whilst the other annotator does not mark the corresponding actions, we still count this as an action alignment mismatch. This creates a harsher measure on action alignment, but gives a better idea of the overall reliability of the annotation task. In Table 8 we show the IAA (as F1) on action alignments. Comparing Tables 8 and 7 we see that, for English at least, the

Lang.	matches	f1
de	300	0.655
en	275	0.769

Table 8: IAA for action alignment, collapsing partial and full alignment

agreement on action alignment is not much lower than that on action identification, indicating that if annotators agree on the actions then they generally agree on how they align. For German, however, the IAA on action alignment is a bit lower, apparently because one of the annotators was much stricter about which actions they aligned.

In order to calculate the IAA on role alignments, we only consider those alignments that connect two roles in aligned frames, of the same type, since these are the only role alignments that count for computing the HMEANT score. This means that if one of the annotators does not align the frames, then all the contained role alignments are counted as mismatches. We do not consider the spans when calculating the agreement on role alignments, meaning that if one annotator has an alignment between roles of type T in frame F, and the other annotator also aligns the same types of roles in the same frame, then they are considered as a match. This is done because it is only the counts of alignments that are relevant for HMEANT scoring. The IAA on the role alignments is quite

Lang.	matches	f1
de	448	0.442
en	506	0.596

Table 9: IAA for role alignment.

low, dipping below 0.5 for German. This is mainly because of the pipelining effect, where annotation disagreements at each stage are compounded. Since the final HMEANT score is computed essentially by counting role alignments, this level of IAA causes problems for this score calculation.

We computed HMEANT and BLEU scores for the hypotheses annotated by each annotator pair. The HMEANT scores were calculated as described in Section 3.2. The two metrics are calculated for each sentence (we apply +1 smoothing for BLEU), then averaged across all sentences. Table 10 shows the scores organised by annotator pair and system type. The agreement in the overall scores is not good, but really just reflects the compounded

Annotator Pair	System	BLEU	HMEANT (Annot. 1)	HMEANT (Annot. 2)
E1, E2	Phrase	0.310	0.626 (2)	0.672 (3)
	Syntax	0.291	0.635 (1)	0.730 (1)
	Rule	0.252	0.578 (3)	0.673 (2)
E1, E3	Phrase	0.378	0.569 (1)	0.602 (3)
	Syntax	0.376	0.553 (2)	0.627 (2)
	Rule	0.320	0.546 (3)	0.646 (1)
E2, E3	Phrase	0.360	0.669 (2)	0.696 (3)
	Syntax	0.362	0.751 (1)	0.739 (1)
	Rule	0.308	0.624 (3)	0.716 (2)
D1, D2	Phrase	0.296	0.327 (1)	0.631 (3)
	Syntax	0.321	0.312 (2)	0.707 (1)
	Rule	0.242	0.274 (3)	0.648 (2)

Table 10: Scores assigned by each annotator pair. The numbers in brackets after the HMEANT scores show the relative ranking assigned by each annotator.

agreement problems in the role alignments (Table 9). In no case do the annotators choose a consistent ranking of the 3 systems, and in 2 of the 4 annotator pairs, the annotators disagree about which is the top performing system.

5.2 Overall Scores

In this section we report the overall HMEANT scores of the three systems whose output we annotated. Our main focus on this paper was on the annotation task, so we do not wish to emphasise the scoring, but it is nevertheless an important end-product of the HMEANT annotation process. The overall scores (HMEANT and +1 smoothed sentence BLEU, averaged across sentences and annotators) are given in Table 11.

Language	System	BLEU	HMEANT
en	Phrase	0.351	0.634
	Syntax	0.344	0.667
	Rule	0.295	0.625
de	Phrase	0.294	0.482
	Syntax	0.302	0.517
	Rule	0.242	0.464

Table 11: Comparison of mean HMEANT and (smoothed sentence) BLEU for the three systems.

From the table we can observe that, whilst BLEU shows similar scores for the phrase-based and syntax-based systems, with lower scores for the rule-based system, HMEANT shows the syntax-based system as being ahead, with the other two showing similar performance. We would caution against reading too much into this, considering the relatively small number of sentences annotated,

and the issues with IAA exposed in the previous section, but it is an encouraging results for syntax-based MT.

5.3 Discussion

Machine translation research needs a reliable method for evaluating and comparing different machine translation systems. The performance of HMEANT as shown in the previous section is disappointing. The fact that the final role IAA, in Table 9, is 0.442 for German and 0.596 for English, demonstrates that there are fundamental problems with the scheme. One of the areas of greatest confusion is between what seems like one of the easiest role types to distinguish: agent and patient. Here is an example of a passive where one annotator has marked "tea" wrongly as agent, and the other annotator correctly labelled it as patient:

Reference: In the kitchen, tea is prepared for the guests
ACTION prepared
LOCATIVE In the kitchen
AGENT / PATIENT tea
MODAL is
BENEFICIARY for the guests

We would argue that the most important change to HMEANT must be in creating more comprehensive annotation guidelines, with examples of difficult cases. Bojar and Wu (2012) listed a number of problems and improvements to HMEANT, which we largely agree with. We list the most important limitations of HMEANT that we have encountered:

- **Single Word Heads** Verbal predicates often consist of multiple words, which can be split. For example: "*Take* him *up* on his offer".
- **Heads being limited to verbs** The semantics of verbs can often be carried by an equivalent noun and should be allowed by HMEANT. For example "My father broke down and cried .", the verb "cried" is correctly paraphrased in "My father collapsed in tears ."
- **Copular Verbs** These do not fit in to the limited list of role types. For example forcing this sentence "The story is plausible", to have and agent and patient is confusing.
- **Prepositional Phrases attaching to a noun** These can greatly affect the semantics of a sentence, but HMEANT has no way of capturing this.
- **Semantics not on head** This frequently occurs with light verbs, for example "Bouson did the review of the paper" is equivalent to "Bouson reviewed the paper".
- **Hierarchy of frames** There are often frames which are embedded in other frames, for example in reported speech. It is not clear whether errors at the lowest level should be marked wrong just at that point, or whether they should be marked wrong all the way up the semantic tree. For example: "Arafat said 'Isreal suffocates such a hope in the germ' ". The frame headed by "said" is largely correct, but the reported speech is not. The patient role of the verb "said" could be aligned as correct, as the error is already captured in relation to the verb "suffocates".
- **No discourse markers** These are important for capturing the relationships between frames and should be labelled.

6 Conclusion

HMEANT represents an attempt to create a human evaluation for machine translation which directly measures the semantic content preserved by the MT. It partly succeeds. However we have cast doubt on the claim that HMEANT can be reliably annotated with minimal annotator training and guidelines. In the most extensive study of inter-annotator agreement yet performed for HMEANT, across two language pairs, we have shown that the disagreements between annotators make it difficult to reliably compare different MT systems with HMEANT scores.

Furthermore, the fact that HMEANT is restricted to annotating purely verbal predicates results in some important disadvantages. Ideally we need a more general definition of a frame, not restricted to purely verbal predicates, and we would like to be able to link frames. We should explore the feasibility of a semantic framework which attempts to overcome reliance on syntactic properties such as Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013).

7 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE).

References

Abend, Omri and Ari Rappoport. 2013. "Universal Conceptual Cognitive Annotation (UCCA)." *Proceedings of ACL*.

Bojar, Ondrej, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. "A Grain of Salt for the WMT Manual Evaluation." *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 1–11. Edinburgh, Scotland.

Bojar, Ondrej and Dekai Wu. 2012. "Towards a Predicate-Argument Evaluation for MT." *Proceedings of SSST*, 30–38.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. "(Meta-) evaluation of machine translation." *Proceedings of the Second Workshop on Statistical Machine Translation*, 136–158. Prague, Czech Republic.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F Zaidan. 2011. "Findings of the 2011 workshop on statistical machine translation." *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 22–64.

Dreyer, Markus and Daniel Marcu. 2012. "Hyter: Meaning-equivalent semantics for translation evaluation." *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 162–171. Montréal, Canada.

Germann, Ulrich. 2008. "Yawat: Yet Another Word Alignment Tool." *Proceedings of the ACL-08: HLT Demo Session*, 20–23. Columbus, Ohio.

Giménez, Jesús and Lluís Màrquez. 2007. "Linguistic features for automatic evaluation of heterogenous mt systems." *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, 256–264. Stroudsburg, PA, USA.

Hutchins, W. J. and H. L. Somers. 1992. *An introduction to machine translation*. Academic Press New York.

Jones, Bevan, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight. 2012. "Semantics-based machine translation with hyperedge replacement grammars." *Proceedings of COLING*.

Koponen, Maarit. 2012. "Comparing human perceptions of post-editing effort with post-editing operations." *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 181–190. Montréal, Canada.

Lavie, Alon and Michael Denkowski. 2009. "The METEOR metric for automatic evaluation of machine translation." *Machine Translation*.

Linguistic Data Consortium. 2002. "Linguistic data annotation specification: Assessment of fluency and adequacy in Chinese-English translation." `http://projects.ldc.upenn.edu/TIDES/Translation/TranAssessSpec.pdf`.

Lo, Chi-kiu, Anand Karthik Tumuluru, and Dekai Wu. 2012. "Fully automatic semantic MT evaluation." *Proceedings of WMT*, 243–252.

Lo, Chi-kiu and Dekai Wu. 2010. "Evaluating machine translation utility via semantic role labels." *Proceedings of LREC*, 2873–2877.

Lo, Chi-kiu and Dekai Wu. 2011a. "MEANT : An inexpensive , high-accuracy , semi-automatic metric for evaluating translation utility via semantic frames." *Proceedings of ACL*, 220–229.

Lo, Chi-kiu and Dekai Wu. 2011b. "Structured vs. flat semantic role representations for machine translation evaluation." *Proceedings of SSST*, 10–20.

Lo, Chi-kiu and Dekai Wu. 2012. "Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics." *Proceedings of SSST*, 49–56.

Lopez, Adam. 2012. "Putting human assessments of machine translation systems in order." *Proceedings of WMT*, 1–9.

NIST. 2005. "The 2005 NIST machine translation evaluation plan (MT-05)." `http://www.itl.nist.gov/iad/mig/tests/mt/2005/doc/mt05_evalplan.v1.1.pdf`.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. "The proposition bank: An annotated corpus of semantic roles." *Computational Linguistics*, 31(1):71–106.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the Association for Computational Linguistics*, 311–318. Philadelphia, USA.

Przybocki, Mark, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. "The NIST

2008 metrics for machine translation challengeoverview, methodology, metrics, and results." *Machine Translation*, 23(2):71–103.

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009a. "Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric." *Proceedings of the Workshop on Statistical Machine Translation at the Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*. Athens, Greece.

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009b. "TER-plus: paraphrase, semantic, and alignment enhancements to translation edit rate." *Machine Translation*.

Tumuluru, Anand Karthik, Chi-kiu Lo, and Dekai Wu. 2012. "Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation." *Proceedings of PACLIC*, 574–581.

Weaver, Warren. 1955. "Translation." William N. Locke and Andrew D. Booth (eds.), *Machine Translation of Languages; Fourteen Essays*, 15–23. Cambridge, MA: MIT Press. Reprint of a memorandum written in 1949.