HUME: Human UCCA-Based Evaluation of Machine Translation

Alexandra Birch[1]*, Omri Abend[2*], Ondřej Bojar[3*], Barry Haddow[1*]
[1]School of Informatics, University of Edinburgh
[2]School of Computer Science and Engineering, Hebrew University of Jerusalem
[3] Charles University in Prague, Faculty of Mathematics and Physics
a.birch@ed.ac.uk, oabend@cs.huji.ac.il
bojar@ufal.mff.cuni.cz, bhaddow@inf.ed.ac.uk

Abstract

Human evaluation of machine translation normally uses sentence-level measures such as relative ranking or adequacy scales. However, these provide no insight into possible errors, and do not scale well with sentence length. We argue for a semantics-based evaluation, which captures what meaning components are retained in the MT output, thus providing a more fine-grained analysis of translation quality, and enabling the construction and tuning of semantics-based MT. We present a novel human semantic evaluation measure, Human UCCA-based MT Evaluation (HUME), building on the UCCA semantic representation scheme. HUME covers a wider range of semantic phenomena than previous methods and does not rely on semantic annotation of the potentially garbled MT output. We experiment with four language pairs, demonstrating HUME's broad applicability, and report good inter-annotator agreement rates and correlation with human adequacy scores.

1 Introduction

Human judgement should be the ultimate test of the quality of an MT system. Nevertheless, common measures for human MT evaluation, such as adequacy and fluency judgements or the relative ranking of possible translations, are problematic in two ways. First, as the quality of translation is multi-faceted, it is difficult to quantify the quality of the entire sentence in a single number. This is indeed reflected in the diminishing inter-annotator agreement (IAA) rates of human ranking measures with the sentence length (Bojar et al., 2011). Second, a sentence-level quality score does not indicate what parts of the sentence are badly translated, and so cannot inform developers in repairing these errors.

These problems are partially addressed by measures that decompose over parts of the evaluated translation, often words or n-grams (see §2 for a brief survey of previous work). A promising line of research decomposes metrics over semantically defined units, quantifying the similarity of the output and the reference in terms of their verb argument structure; the most notable of these measures is HMEANT (Lo and Wu, 2011).

We propose the HUME metric, a human evaluation measure that decomposes over UCCA semantic units. UCCA (Abend and Rappoport, 2013) is an appealing candidate for semantic analysis, due to its cross-linguistic applicability, support for rapid annotation, and coverage of many fundamental semantic phenomena, such as verbal, nominal and adjectival argument structures and their inter-relations.

HUME operates by aggregating human assessments of the translation quality of individual semantic units in the source sentence. We are thus avoiding the semantic annotation of machine-generated text, which is often garbled or semantically unclear. This also allows the re-use of the source semantic annotation for measuring the quality of different translations of the same source sentence and avoids relying on reference translations, which have been shown to bias annotators (Fomicheva and Specia, 2016).

After a brief review (§2), we describe HUME in

* All authors contributed equally to this work.

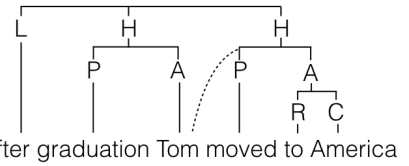
Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1264–1274,
Austin, Texas, November 1-5, 2016.

detail (§3). Our experiments with four language pairs: English to Czech, German, Polish and Romanian (§4) document HUME's inter-annotator agreement and efficiency (time of annotation). We further empirically compare HUME with direct assessment of human adequacy ratings (§5), and conclude by discussing the differences with HMEANT (§6).

2 Background

MT Evaluation. Human evaluation is generally done by ranking the outputs of multiple systems e.g., in the WMT tasks (Bojar et al., 2015), or by assigning adequacy/fluency scores to each translation, a procedure recently improved by Graham et al. (2015b) under the title Direct Assessment. We use this latter method to compare and contrast with HUME later in the paper. HTER (Snover et al., 2006) is another widely used human evaluation metric which uses edit distance metrics to compare a translation and its human post-edition. HTER suffers from the problem that small edits in the translation could in fact be serious flaws in accuracy, e.g., deleting a negation. Some manual measures ask annotators to explicitly mark errors, but this has been found to have even lower agreement than ranking (Lommel et al., 2014).

However, while providing the gold standard for MT evaluation, human evaluation is not a scalable solution. Scalability is addressed by employing automatic and semi-automatic approximations of human judgements. Commonly, such scores decompose over the sub-parts of the translation, and quantify how many of these sub-parts appear in a manually created reference translation. This decomposition allows system developers to localize the errors. The most commonly used measures decompose over n-grams or individual words, e.g., BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and METEOR (Banerjee and Lavie, 2005). Another common approach is to determine the similarity between the reference and translation in terms of string edits (Snover et al., 2006). While these measures stimulated much progress in MT research by allowing the evaluation of massive-scale experiments, the focus on words and n-grams does not provide a good estimate of semantic correctness, and may favour shallow string-based MT models.



L	Linker	A	Participant
H	Parallel Scene	R	Relater
P	Process	C	Centre

Figure 1: Sample UCCA annotation. Leaves correspond to words and nodes to units. The dashed edge indicates that "Tom" is also a participant in the "moved to America" Scene. Edge labels mark UCCA categories.

In order to address this shortcoming, more recent work quantified the similarity of the reference and translation in terms of their structure. Liu and Gildea (2005) took a syntactic approach, using dependency grammar, and Owczarzak et al. (2007) took a similar approach using Lexical Functional Grammar structures. Giménez and Màrquez (2007) proposed to combine multiple types of information, capturing the overlap between the translation and reference in terms of their semantic (predicate-argument structures), lexical and morphosyntactic features. Macháček and Bojar (2015) divided the source sentences into shorter segments, defined using a phrase structure parse, and applied human ranking to the resulting segments.

Perhaps the most notable attempt at semantic MT evaluation is MEANT and its human variant HMEANT (Lo and Wu, 2011), which quantifies the similarity between the reference and translation in terms of the overlap in their verbal argument structures and associated semantic roles. We discuss the differences between HMEANT and HUME in §6.

Semantic Representation. UCCA (Universal Conceptual Cognitive Annotation) (Abend and Rappoport, 2013) is a cross-linguistically applicable scheme for semantic annotation. Formally, an UCCA structure is a directed acyclic graph (DAG), whose leaves correspond to the words of the text. The graph's nodes, called *units*, are either terminals or several elements jointly viewed as a single entity according to some semantic or cognitive consideration. Edges bear a category, indicating the role of the sub-unit in the structure the unit represents.

UCCA's basic inventory of distinctions (its *foundational layer*) focuses on argument structures (ad-

jectival, nominal, verbal and others) and relations between them. The most basic notion is the *Scene*, which describes a movement, an action or a state which persists in time. Each Scene contains one main relation and zero or more participants. For example, the sentence "After graduation, Tom moved to America" contains two Scenes, whose main relations are "graduation" and "moved". The participant "Tom" is a part of both Scenes, while "America" only of the latter (Figure 1). Further categories account for inter-scene relations and the sub-structures of participants and relations.

The use of UCCA for semantic MT evaluation has several motivations. First, UCCA's foundational layer can be annotated by non-experts after a short training (Abend and Rappoport, 2013; Marinotti, 2014). Second, UCCA is cross-linguistically applicable, seeking to represent what is shared between languages by building on linguistic typological theory (Dixon, 2010b; Dixon, 2010a; Dixon, 2012). Its cross-linguistic applicability has so far been tested in annotations of English, French, German and Czech. Third, the scheme has been shown to be stable across translations: UCCA annotations of translated text usually contain the same set of relations (Sulem et al., 2015), indicating that UCCA reflects a layer of representation that in a correct translation is mostly shared between the translation and the source.

The Abstract Meaning Representation (AMR) (Banarescu et al., 2013) shares UCCA's motivation for defining a more complete semantic annotation. However, using AMR is not optimal for defining a decomposition of a sentence into semantic units as it does not anchor its semantic symbols in the text, and thus does not provide clear decomposition of the sentence into sub-spans. Also, AMR is more fine-grained than UCCA and consequently harder to annotate. Other approaches represent semantic structures as bi-lexical dependencies (Sgall et al., 1986; Hajič et al., 2012; Oepen and Lønning, 2006), which are indeed anchored in the text, but are less suitable for MT evaluation as they require linguistic expertise for their annotation.

3 The HUME Measure

3.1 Annotation Procedure

This section summarises the manual annotation procedure used to compute the HUME measure. We denote the source sentence as s and the translation as t. The procedure involves two manual steps: (1) UCCA-annotating s, (2) HUME-annotation: human judgements as to the translation quality of each semantic unit of s relative to t, where units are defined according to the UCCA annotation. UCCA annotation is performed once for every source sentence, irrespective of the number of its translations we wish to evaluate, and requires proficiency in the source language only. HUME annotation requires the employment of bilingual annotators.[1]

UCCA Annotation. We begin by creating UCCA annotations for the source sentence, following the UCCA guidelines.[2] A UCCA annotation for a sentence s is a labeled DAG G, whose leaves are the words of s. For every node in G, we define its *yield* to be its leaf descendants. The semantic units for s according to G are the yields of nodes in G.

Translation Evaluation. HUME annotation is done by traversing the semantic units of the source sentence, which correspond to the arguments and relations expressed in the text, and marking the extent to which they have been correctly translated. HUME aggregates the judgements of the users into a composite score, which reflects the overall extent to which the semantic content of s is preserved in t.

Annotation of the semantic units requires first deciding whether a unit is *structural*, i.e., has meaning-bearing sub-units in the target language, or *atomic*. In most cases, atomic units correspond to individual words, but they may also correspond to multi-word expressions that translate as one unit. For instance, the expression "took a shower" is translated into the German "duschte", while its individual words do not correspond to any sub-part of the German translation, motivating the labeling the entire expression as an atomic node. When a multi-word unit is labeled

[1] Where bilingual annotators are not available, the evaluation could be based on the UCCA structure for the *reference* translation. See discussion in §6.

[2] All UCCA-related resources can be found here: `http://www.cs.huji.ac.il/~oabend/ucca.html`

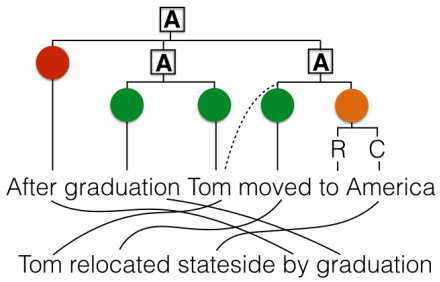


Figure 2: HUME annotation of an UCCA tree with a word-aligned example translation shown below. Atomic units are labelled using traffic lights (Red, Orange, Green) and structural units are marked A or B.

as atomic, its sub-units' annotations are ignored in the evaluation.

Atomic units can be labelled as "Green" (G, correct), "Orange" (O, partially correct) and "Red" (R, incorrect). Green means that the meaning of the word or phrase has been largely preserved. Orange means that the essential meaning of the unit has been preserved, but some part of the translation is wrong. This is often be due to the translated word having the wrong inflection, in a way that impacts little on the understandability of the sentence. Red means that the essential meaning of the unit has not been captured.

Structural units have sub-units (children in the UCCA graph), which are themselves atomic or structural. Structural units are labeled as "Adequate" (A) or "Bad" (B), meaning that the relation between the sub-units went wrong[3]. We will use the example "man bites dog" to illustrate typical examples of why a structural node should be labelled as "Bad": incorrect ordering ("dog bites man"), deletion ("man bites") and insertion ("man bites biscuit dog").

HUME labels reflect adequacy, rather than fluency judgements. Specifically, annotators are instructed to label a unit as Adequate if its translation is understandable and preserves the meaning of the source unit, even if its fluency is impaired.

Figure 2 presents an example of a HUME annotation, where the translation is in English for ease of comprehension. When evaluating "to America" the annotator looks at the translation and sees the word "stateside". This word captures the whole phrase and so we mark this non-leaf node with an atomic label. Here we choose Orange since it approximately captures the meaning in this context. The ability to mark non-leaves with atomic labels allows the annotator to account for translations which only correspond at the phrase level. Another feature highlighted in this example is that by separating structural and atomic units, we are able to define where an error occurs, and localise the error to its point of origin. The linker "After" is translated incorrectly as "by" which changes the meaning of the entire sentence. This error is captured at the atomic level, and it is labelled Red. The sentence still contains two Scenes and a Linker and therefore we mark the root node as structurally correct, Adequate.

3.2 Composite Score

We proceed to detailing how judgements on the semantic units of the source are aggregated into a composite score. We start by taking a very simple approach and compute an accuracy score. Let $\mathit{Green}(s,t)$, $\mathit{Adequate}(s,t)$ and $\mathit{Orange}(s,t)$ be the number of Green, Adequate and Orange units, respectively. Let $\mathit{Units}(s)$ be the number of units marked with any of the labels. Then HUME's composite score is:

$$\mathit{HUME}(s,t) = \frac{\mathit{Green}(s,t) + \mathit{Adequate}(s,t) + 0.5 \cdot \mathit{Orange}(s,t)}{\mathit{Units}(s)}$$

3.3 Annotation Interface

Figure 3 shows the HUME annotation interface[4]. One source sentence and one translation are presented at a time. The user is asked to select a label for each source semantic unit, by clicking the "A", "B", Green, Orange, or Red buttons to the right of the unit's box. Units with multiple parents (as with "Tom" in Figure 2) are displayed twice, once under each of their parents, but are only annotatable in one of their instances, to avoid double counting.

The interface presents, for each unit, the translation segment aligned with it. This allows the user, especially in long sentences, to focus her attention on the parts that are most likely to be relevant for her judgement. As the alignments are automatically derived, and therefore noisy, the annotator is instructed to treat the aligned text is a cue, but to ignore the alignment if it is misleading, and instead make a

[3]Three labels are used with atomic units, as opposed to two labels with structural units, as atomic units are more susceptible to slight errors.

[4]A demo of HUME can be found in `www.cs.huji.ac.il/~oabend/hume_demo.html`

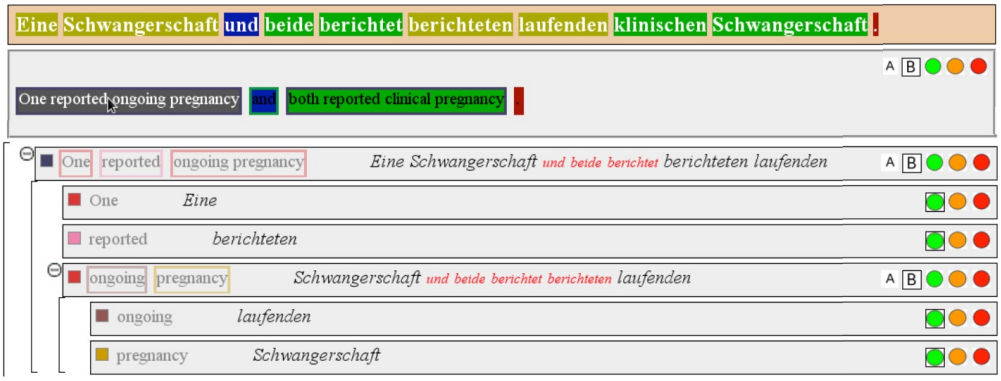


Figure 3: The HUME annotation tool. The top orange box contains the translation. The source sentence is directly below it, followed by the tree of the source semantic units. Alignments between the source and translation are in italics and unaligned intervening words are in red (see text).

judgement according to the full translation. Concretely, let s be a source sentence, t a translation, and $A \subset 2^s \times 2^t$ a many-to-many word alignment. If u is a semantic unit in s, whose yield is $yld(u)$, we define the aligned text in t to be $\bigcup_{(x_s,x_t)\in A \wedge x_s \cap yld(u) \neq \emptyset} x_t$.

Where the aligned text is discontinuous in t, words between the left and right boundaries which are not contained in it (intervening words) are presented in a smaller red font. Intervening words are likely to change the meaning of the translation of u, and thus should be attended to when considering whether the translation is correct or not.

For example, in Figure 3, "ongoing pregnancy" is translated to "Schwangerschaft ... laufenden" (lit. "pregnancy ... ongoing"). This alone seems acceptable but the interleaving words in red notify the annotator to check the whole translation, in which the meaning of the expression is not preserved[5]. The annotator should thus mark this structural node as Bad.

4 Experiments

In order to validate the HUME metric, we ran an annotation experiment with one source language (English), and four target languages (Czech, German, Polish and Romanian), using text from the public health domain. Semantically accurate translation is paramount in this domain, which makes it particularly suitable for semantic MT evaluation. HUME is evaluated in terms of its consistency (inter-annotator agreement), efficiency (time of annotation) and validity (by comparing it with crowd-sourced adequacy judgements).

4.1 Datasets and Translation Systems

For each of the four language pairs under consideration we built phrase-based MT systems using Moses (Koehn et al., 2007). These were trained on large parallel data sets extracted from OPUS (Tiedemann, 2009), and the data sets released for the WMT14 medical translation task (Bojar et al., 2014), giving between 45 and 85 million sentences of training data, depending on the language pair. These translation systems were used to translate texts derived from both NHS 24[6] and Cochrane[7] into the four languages. NHS 24 is a public body providing healthcare and health-service related information in Scotland; Cochrane is an international NGO which provides independent systematic reviews on health-related research. NHS 24 texts come from the "Health A-Z" section in the NHS Inform website, and Cochrane texts come from their plain language summaries and abstracts.

4.2 HUME Annotation Statistics

The source sentences are all in English, and their UCCA annotation was performed by four computational linguists and one linguist. For the annotation of the MT output, we recruited two annotators for each of German, Romanian and Polish and one main annotator for Czech. For computing Czech IAA, several further annotators worked on a small number of sentences each. We treat these further annotators

[5]The interleaving words are "... und beide berichtet berichteten ..." (lit. "... and both report reported ..."), which doesn't form any coherent relation with the rest of the sentence.

[6]http://www.nhs24.com/

[7]http://www.cochrane.org/

		cs	de	pl	ro
#Sentences	Annot. 1	324	339	351	230
	Annot. 2	205	104	340	337
#Units	Annot. 1	8794	9253	9557	6152
	Annot. 2	5553	2906	9303	9228

Table 1: HUME-annotated #sentences and #units.

	cs	de	pl	ro
Annot. 1	255	140	138	96
Annot. 2	*	162	229	207

Table 2: Median annotation times per sentence, in seconds. *: no timing information is available, as this was a collection of annotators, working in parallel.

as one annotator, resulting in two annotators for each language pair. The annotators were all native speakers of the respective target languages and fluent in English. They completed a three hour on-line training session which included a description of UCCA and the HUME task, followed by walking through a few examples.

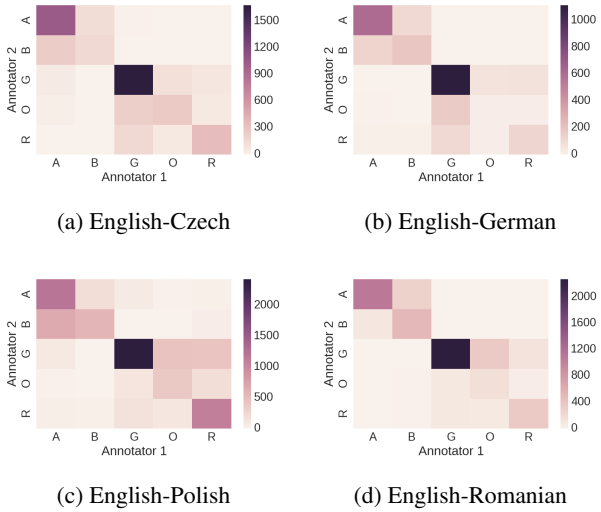
Table 1 shows the total number of sentences and units annotated by each annotator. Not all units in all sentences were annotated, often due to the annotator accidentally missing a node.

Efficiency. We estimate the annotation time using the timestamps provided by the annotation tool, which are recorded whenever an annotated sentence is submitted. Annotators are not able to re-open a sentence once submitted. To estimate the annotation time, we compute the time difference between successive sentences, and discard outlying times, assuming annotation was not continuous in these cases. From inspection of histograms of annotation times, we set the upper threshold at 500 seconds. Median annotation times are presented in Table 2, indicating that the annotation of a sentence takes around 2–4 minutes, with some variation between annotators.

Inter-Annotator Agreement. In order to assess the consistency of the annotation, we measure the Inter-Annotator Agreement (IAA) using Cohen's Kappa on the multiply-annotated units. Table 3 reports the number of units which have two annotations from different annotators and the corresponding Kappas. We report the overall Kappa, as well as separate Kappas on atomic units (annotated as Red, Orange or Green) and structural units (annotated as

	cs	de	pl	ro
Sentences	181	102	334	217
All units	4686	2793	8384	5604
Kappa	0.64	0.61	0.58	0.69
Atomic units	2982	1724	5386	3570
Kappa	0.54	0.29	0.54	0.50
Structural units	1602	1040	2655	1989
Kappa	0.31	0.44	0.33	0.58

Table 3: IAA for the multiply-annotated units, measured by Cohen's Kappa.



(a) English-Czech (b) English-German

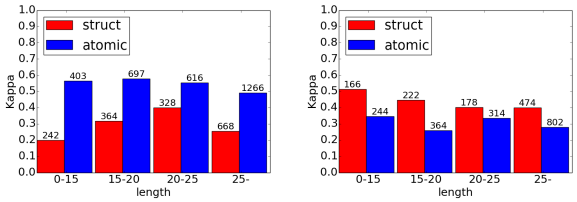
(c) English-Polish (d) English-Romanian

Figure 4: Confusion matrices for each language pair.

Adequate or Bad). As expected and confirmed by confusion matrices in Figure 4, there is generally little confusion between the two types of units. This results in the Kappa for all units being considerably higher than the Kappa over the atomic units or structural units, where there is more internal confusion.

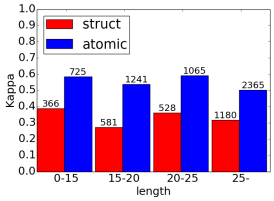
To assess HUME reliability for long sentences, we binned the sentences according to length and measured Kappa on each bin (Figure 5). We see no discernible reduction of IAA with sentence length. Table 3 also shows that the overall IAA is similar for all languages, presenting good agreement (0.6–0.7). However, there are differences observed when we break down by node type. Specifically, we see a contrast between Czech and Polish, where the IAA is higher for atomic than for structural units, and German and Romanian, where the reverse is true. We also observe low IAA (around 0.3) in the cases of German atomic units, and Polish and Czech structural units.

Looking more closely at the areas of disagree-

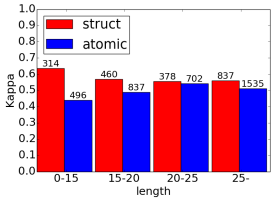


(a) English-Czech

(b) English-German



(c) English-Polish



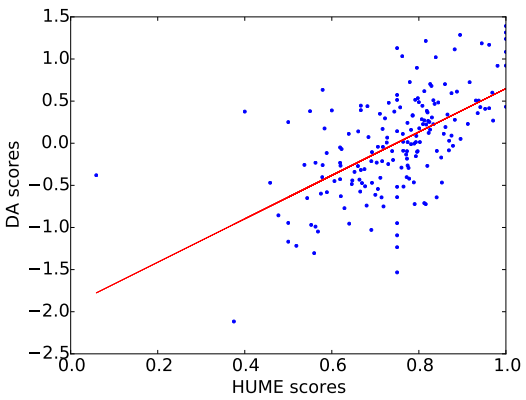
(d) English-Romanian

Figure 5: Kappa versus sentence length for structural and atomic units. (Node counts in bins on top of each bar.)

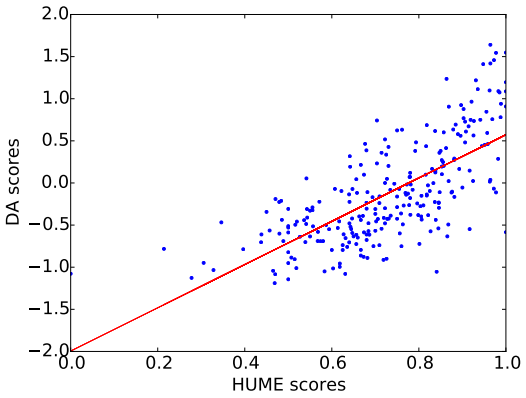
ment, we see that for the Polish structural units, the proportion of As was quite different between the two annotators (53% vs. 71%), whereas for other languages the annotators agree in the proportions. We believe that this was because one of the Polish annotators did not fully understand the guidelines for structural units, and percolated errors up the tree, creating more Bs. For German atomic and Czech structural units, where Kappa is also around 0.3, the proportion of such units being marked as "correct" is relatively high, meaning that the class distribution is more skewed, so the expected agreement used in the Kappa calculation is high, lowering Kappa. Finally we note some evidence of domain-specific disagreements, for instance the German MT system normally translated "review" (as in "systematic review" – a frequent term in the Cochrane texts) as "Überprüfung", which one annotator marked correct, and the other (a Cochrane employee) as incorrect.

5 Comparison with Direct Assessment

Recent research (Graham et al., 2015b; Graham et al., 2015a; Graham, 2015) has proposed a new approach for collecting accuracy ratings, direct assessment (DA). Statistical interpretation of a large number of crowd-sourced adequacy judgements for each candidate translation on a fine-grained scale of 0 to 100 results in reliable aggregate scores, that correlate very strongly with one another.



(a) English-German



(b) English-Romanian

Figure 6: HUME vs DA scores. DA score have been standardised for each crowdsourcing annotator and averaged across exactly 10 annotators. HUME scores are averaged where there were two annotations.

We attempted to follow Graham et al. (2015b) but struggled to get enough crowd-sourced judgements for our target languages. We ended up with 10 adequacy judgements on most of the HUME annotated translations for German and Romanian but insufficient data for Czech and Polish. We see this as a severe practical limitation of DA.

Figure 6 plots the HUME score for each sentence against its DA score. HUME and Direct Assessment scores correlate reasonably well. The Pearson correlation for en-ro (en-de) is 0.70 (0.58), or 0.78 (0.74) if only doubly HUME-annotated points are considered. This confirms that HUME is consistent with an accepted human evaluation method, despite their conceptual differences. While DA is a valuable tool, HUME has two advantages: it returns fine-grained

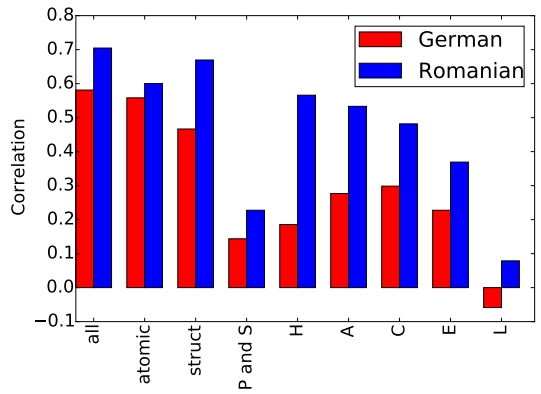


Figure 7: Pearson correlation of HUME vs. DA scores for en-ro and en-de. Each bar represents a correlation between DA and an aggregate HUME score based on a sub-set of the units (#nodes for the en-de/en-ro setting in brackets): all units ("all", 8624/10885), atomic ("atomic", 5417/6888) and structural units ("struct", 3207/3997), and units by UCCA categories: Scene main relations (i.e, Process and State units; "P and S", 954/1178), Parallel Scenes ("H", 656/784), Participants ("A", 1348/1746), Centres ("C", 1904/2474), elaborators ("E", 1608/2031) and linkers ("L", 261/315).

semantic information about the quality of translations and it only requires very few annotators. Direct assessment returns a single opaque score, and (as also noted by Graham et al.) requires a large crowd which may not be available or reliable.

Figure 7 presents an analysis of HUME's correlations with DA by HUME unit type, an analysis enabled by HUME's semantic decomposition. For both target languages, correlation is highest in the 'all' case, supporting our claim for the value of aggregating over a wide range of semantic phenomena. Some types of nodes predict the DA scores better than others. HUME scores on As correlate more strongly with DA than scores on Scene Main Relations (P+S). Center nodes (C) are also more correlated than elaborator nodes (E), which is expected given that Centers are defined to be more semantically dominant. Future work will construct an aggregate HUME score which weights the different node types according to their semantic prominence.

HUME and DA are conceptually very different metrics: while DA standardises and averages scores across annotators to denoise the crowd-sourced raw data, thus obtaining a single aggregate score, HUME decomposes over a combinatorial structure, thus allowing to localize the translation errors. We now turn to comparing HUME to a more conceptually-related measure, namely HMEANT.

6 Comparison with HMEANT

HMEANT is a human MT evaluation metric that measures the overlap between the translation a reference in terms of their SRL annotations. In this section we present a qualitative comparison between HUME and HMEANT, using examples from our experimental data.

Verbal Structures Only? HMEANT focuses on verbal argument structures, ignoring other pervasive phenomena such as non-verbal predicates and interclausal relations. Consider the following example:

Source	a coronary angioplasty may not be technically possible
Transl.	eine koronare Angioplastie kann nicht technisch möglich
Gloss	a coronary angioplasty can not technically possible

The German translation is largely correct, except that the main verb "sein" ("be") is omitted. While this may be interpreted as a minor error, HMEANT will assign the sentence a very low score, as it failed to translate the main verb.

It is also relatively common that verbal constructions are translated as non-verbal ones or vice versa. Consider the following example:

Source	... tend to be higher in saturated fats
Transl.	... in der Regel höher in gesättigte Fette
Gloss	... as a rule higher in saturated fats

The German translation is largely correct, despite the grammatical divergence, namely that the English verb "tend" is translated into the German prepositional phrase "in der Regel" ("as a rule"). HMEANT will consider the translation to be of poor quality as there is no German verb to align with the English one.

We conducted an analysis of the English UCCA Wikipedia corpus (5324 sentences) in order to assess the pervasiveness of three phenomena that are not well supported by HMEANT.[8] First, copula clauses

[8] Argument structures and linkers are explicitly marked in UCCA. Non-auxiliary instances of "be" and nouns are identi-

are treated in HMEANT simply as instances of the main verb "be", which generally does not convey the meaning of these clauses. They appear in 21.7% of the sentences, according to conservative estimates that only consider non-auxiliary instances of "be". Second, nominal argument structures, ignored by HMEANT, are in fact highly pervasive, appearing in 48.7% of the sentences. Third, linkers that express inter-relations between clauses (mainly discourse markers and conjunctions) appear in 56% of the sentences, but are again ignored by HMEANT. For instance, linkers are sometimes omitted in translation, but these omissions are not penalized by HMEANT. The following is such an example from our experimental dataset:

Source	However, this review was restricted to ...
Transl.	Diese Überprüfung bescränkte sich auf ...
Gloss	This review was restricted to ...

We note that some of these issues were already observed in previous applications of HMEANT to languages other than English. See Birch et al. (2013) for German, Bojar and Wu (2012) for Czech and Chuchunkov et al. (2014) for Russian.

One Structure or Two. HUME only annotates the source, while HMEANT relies on two independently constructed structural annotations, one for the reference and one for the translation. Not annotating the translation is appealing as it is often impossible to assign a semantic structure to a low quality translation. On the other hand, HUME may be artificially boosting the perceived understandability of the translation by allowing access to the source.

Alignment. In HMEANT, the alignment between the reference and translation structures is a key part of the manual annotation. If the alignment cannot be created, the translation is heavily penalized. Bojar and Wu (2012) and Chuchunkov et al. (2014) argue that the structures of the reference and of an accurate translation may still diverge, for instance due to a different interpretation of a PP-attachment, or the verb having an additional modifier in one of the structures. It would be desirable to allow modifications to the SRL annotations at the alignment stage, to avoid unduly penalizing such spurious divergences.

The same issue is noted by Lo and Wu (2014): the IAA on SRL dropped from 90% to 61% when the two aligned structures were from two different annotators. HUME uses automatic (word-level) alignment, which only serves as a cue for directing the attention of the annotators. The user is expected to mentally correct the alignment as needed, thus circumventing this difficulty.

Monolingual vs. Bilingual Evaluation. HUME diverges from HMEANT and from shallower measures like BLEU, in not requiring a reference. Instead, it directly compares the source and the translation. This requires the employment of bilingual annotators, but has the benefit of avoiding using a reference, which is never uniquely defined, and may thus lead to unjustly low scores where the translation is a paraphrase of the reference. If only monolingual annotators are available, the HUME evaluation could be performed with a reference sentence instead of with the source. This, however, would risk inaccurate judgements due to the naturally occurring differences between the source and its reference translations.

7 Conclusion

We have introduced HUME, a human semantic MT evaluation measure which addresses a wide range of semantic phenomena. We have shown that it can be reliably and efficiently annotated in multiple languages, and that annotation quality is robust to sentence length. Comparison to direct assessments further support HUME's validity. We believe that HUME, and a future automated version of HUME, allows for a finer-grained analysis of translation quality, and will be useful in informing the development of a more semantically aware approach to MT.

All annotation data gathered in this project, together with analysis scripts, is available online[9].

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 644402 (HimL).

fied using the NLTK standard tagger. Nominal argument structures are here Scenes whose Main Relation is headed by a noun.

[9] https://github.com/bhaddow/hume-emnlp16

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria, August. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI, USA. Association for Computational Linguistics.

Alexandra Birch, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. 2013. The feasibility of HMEANT as a human MT evaluation metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 52–61, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar and Dekai Wu. 2012. Towards a Predicate-Argument Evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 30–38, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar F. Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Alexander Chuchunkov, Alexander Tarelkin, and Irina Galinskaya. 2014. Applying HMEANT to English-Russian Translations. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 43–50, Doha, Qatar, October. Association for Computational Linguistics.

Robert M.W. Dixon. 2010a. *Basic Linguistic Theory: Grammatical Topics*, volume 2. Oxford University Press.

Robert M.W. Dixon. 2010b. *Basic Linguistic Theory: Methodology*, volume 1. Oxford University Press.

Robert M.W. Dixon. 2012. *Basic Linguistic Theory: Further Grammatical Topics*, volume 3. Oxford University Press.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Diego, CA, USA. Morgan Kaufmann Publishers Inc.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *54th Annual Meeting of the Association for Computational Linguistics*, ACL, Berlin, Germany.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015a. Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, pages 1–28.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015b. Accurate evaluation of segment-level machine translation metrics. In *Proc. of NAACL-HLT*, pages 1183–1191.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China, July. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In

Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12), pages 3153–3160, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Chi-kiu Lo and Dekai Wu. 2011. Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 10–20. Association for Computational Linguistics.

Chi-Kiu Lo and Dekai Wu. 2014. On the Reliability and Inter-Annotator Agreement of Human Semantic MT Evaluation via HMEANT. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arle Richard Lommel, Maja Popovic, and Aljoscha Burchardt. 2014. Assessing Inter-Annotator Agreement for Translation Error Annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*. LREC.

Matouš Macháček and Ondřej Bojar. 2015. Evaluating Machine Translation Quality Using Short Segments Annotations. *The Prague Bulletin of Mathematical Linguistics*, 103:85–110, April.

Pedro Marinotti. 2014. Measuring semantic preservation in machine translation with HCOMET: human cognitive metric for evaluating translation. Master's thesis, University of Edinburgh.

Stephan Oepen and Jan Tore Lønning. 2006. Discriminant-based MRS banking. In *Proceedings of LREC*, pages 1250–1255.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Evaluating machine translation with LFG dependencies. *Machine Translation*, 21(2):95–119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA. Association for Computational Linguistics.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Elior Sulem, Omri Abend, and Ari Rappoport. 2015. Conceptual annotations preserve structure across translations: A French-English case study. In *ACL 2015 Workshop on Semantics-Driven Statistical Machine Translation (S2MT)*, pages 11–22.

Jörg Tiedemann. 2009. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, Borovets, Bulgaria. John Benjamins.