Towards a Predicate-Argument Evaluation for MT

Ondřej Bojar[α], **Dekai Wu**[β]
[α] Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
[β] *HKUST*, Human Language Technology Center,
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
bojar@ufal.mff.cuni.cz, dekai@cs.ust.hk

Abstract

HMEANT (Lo and Wu, 2011a) is a manual MT evaluation technique that focuses on predicate-argument structure of the sentence. We relate HMEANT to an established linguistic theory, highlighting the possibilities of reusing existing knowledge and resources for interpreting and automating HMEANT. We apply HMEANT to a new language, Czech in particular, by evaluating a set of English-to-Czech MT systems. HMEANT proves to correlate with manual rankings at the sentence level better than a range of automatic metrics. However, the main contribution of this paper is the identification of several issues of HMEANT annotation and our proposal on how to resolve them.

1 Introduction

Manual evaluation of machine translation output is a tricky enterprise. It has been long recognized that different evaluation techniques lead to different outcomes, e.g. Blanchon et al. (2004) mention an evaluation carried out in 1972 where the very same Russian-to-English MT outputs were scored 4.5 out of the maximum 5 points by prospective users of the system but only 1 out of 5 by teachers of English. Throughout the years, many techniques were explored with more or less of a success.

The two-scale scoring for adequacy and fluency used in NIST evaluation has been abandoned by some evaluation campaigns, most notably the WMT shared task series, see Koehn and Monz (2006) through Callison-Burch et al. (2012)[1]. Since 2008, WMT uses a simple relative ranking of MT outputs as its primary manual evaluation technique: the annotator is presented with up to 5 MT outputs for a given input sentence and the task is to rank them from best to worst (ties allowed) on whatever criteria he or she deems appropriate. While this single-scale relative ranking is perhaps faster to annotate and reaches a higher inter- and intra-annotator agreement than the (absolute) fluency and adequacy (Callison-Burch et al., 2007), the technique and its evaluation are still far from satisfactory. Bojar et al. (2011) observe several discrepancies in the interpretation of the rankings, partly due to the high load on human annotators (the comparison of several long sentences at once, among other issues) but partly also due to technicalities of the calculation.

Lo and Wu (2011a) present an interesting evaluation technique called MEANT (or HMEANT if carried out by humans), the core of which lies in assessing whether the key elements in the predicate-argument structure of the sentence have been preserved. In other words, lay annotators are checking, if they recognize *who did what* to *whom*, *when*, *where* and *why* from the MT outputs and whether the respective role fillers convey the same meaning as in the reference translation. HMEANT has been shown to correlate reasonably well with manual adequacy and ranking evaluations. It is relatively fast and should lend itself to full automatization. On the other hand, HMEANT was so far tested only on translation into English and with just three competing MT systems.

[1] http://www.statmt.org/wmt06 till wmt12

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 30–38,
Jeju, Republic of Korea, 12 July 2012.

In this work, we extend the application of HMEANT to evaluating MT into Czech, a morphologically rich language with relatively free word order. The paper is structured as follows: Section 2 presents the technical details of HMEANT and relates HMEANT to an established linguistic theory that underlies the Prague dependency treebanks (Hajič et al., 2006; Hajič et al., 2012) and several other works. We also suggest possible benefits of this coupling such as the reuse of tools. In Section 3, we describe the setup and results of our HMEANT experiment. Since this is the first time HMEANT is applied to a new language, Section 4 constitutes the main contribution of this work. We point out at several problems of HMEANT and propose a remedy, the empirical evaluation of which however remains for future work. Section 5 concludes our observations.

2 Relating HMEANT and Valency Theory of FGD

2.1 HMEANT Annotation Procedure

HMEANT is designed to be simple and fast. The annotation consists of two steps: (1) semantic role labelling, SRL in the sequel, and (2) alignment of roles between the hypothesis and the reference.

The annotation guidelines are deliberately minimalistic, so that even inexpert people can learn them quickly. The complete guidelines for SRL are given in Figure 1 and it takes less than 15 minutes to train an unskilled person.

In the alignment task, the annotators first indicate which frames in the reference and the hypothesis correspond to each other. In the second step, they align all matching role fillers to each other and also mark the translation as "Correct" or "Partial".

The HMEANT calculation then evaluates the f-score of the predicates and their role fillers in a given sentence. An important aspect of the calculation is that unmatched predicates with all their role fillers are excluded from the calculation.

2.2 Functional Generative Description

The core ideas of HMEANT follow the case grammar (Fillmore, 1968) or PropBank (Palmer et al., 2005) and can be also directly related to an established linguistic theory which was primarily developed for Czech, namely the Functional Generative Description (Sgall, 1967; Sgall et al., 1986). The theory defines so-called "tectogrammatical" layer (t-layer). At the t-layer, each sentence is represented as a dependency tree with just content words as separate nodes. All auxiliary words are "hidden" into attributes of the corresponding t-nodes. Moreover, ellipsis is restored to some extent, so e.g. dropped subject pronouns do have a corresponding t-node.

Semantic frames summarize a sentence using a simple event structure that captures essential parts of the meaning like "who did what to whom, when, where, why and how".
Phrases or clauses that express meanings can be identified as playing a particular semantic role in the sentence. In other words, semantic frames are the systematic abstraction of the meanings in a sentence.
The following is the list of the semantic roles to be used in HMEANT evaluation:

Agent (who)	Action (did)
Experiencer or Patient (what)	Benefactive (whom)
Temporal (when)	Locative (where)
Purpose (why)	Manner (how)
Degree or Extent (how)	Modal (how) [may, should, ...]
Negation (how) [not]	Other adverbial argument (how)

You may consider the Action predicate to be the central event, while the other roles modify the Action to give a more detailed description of the event. Each semantic frame contains exactly one Action and any number of other roles.
Please note that the Action predicate must be exactly ONE single word.
There may be multiple semantic frames in one sentence, because a sentence may be constructed to describe multiple events and each semantic frame captures only one event.

Figure 1: Semantic role labeling guidelines of HMEANT.

An important element of FGD is the valency theory (Panevová, 1980) which introduces empirical linguistic tests to distinguish between what other theories would call complements vs. adjuncts and postulates the relationship between the set of verb modifiers as observed in the sentence and the set of valency slots that should be listed in a valency dictionary. This aspect could provide a further refinement of HMEANT, e.g. weighing complements and adjuncts differently.

FGD has been thoroughly tested and refined during the development of the Prague Dependency Treebank (Hajič et al., 2006)[2] and the parallel Prague Czech-English Dependency Treebank (Hajič

[2] http://ufal.mff.cuni.cz/pdt2.0/

et al., 2012)[3]. Note that the latter is a translation of all the 49k sentences of the Penn Treebank WSJ section. Both English and Czech sentences are manually annotated at the tectogrammatical layer, where the English layer is based on the Penn annotation and manually adapted for t-layer. Both languages include their respective valency lexicons and the work on a bilingual valency lexicon is being developed (Šindlerová and Bojar, 2010).

A range of automatic tools to convert plain text up to the t-layer exist for both English and Czech. Most of them are now part of the Treex platform (Popel and Žabokrtský, 2010)[4] and they were successfully used in automatic annotation of 15 million parallel sentences (Bojar et al., 2012)[5] as well as other NLP tasks including English-to-Czech MT. Recently, significant effort was also invested in parsing not quite correct output of MT systems into Czech for the purposes of rule-based grammar correction (Rosa et al., 2012). Establishing the automatic pipeline for MEANT should be relatively easy with these tools at hand.

2.3 HMEANT vs. FGD Valency

The formulation of HMEANT in terms of FGD is straightforward: it is the f-score of matched t-nodes for predicates and the subtrees of their immediate dependents in the t-trees of the hypothesis and the reference.

HMEANT uses a simple web-based annotation interface which operates on the surface form of the sentence. Annotators mark the predicate and their complementations as contiguous spans in the sentence. While this seems natural when we want lay people to annotate, it brings some problems, see Section 4. A linguistically adequate interface would allow to mark tectogrammatical nodes and subtrees in the t-layer, however, the customizable editor TrEd[6] used for manual annotation of t-layer is too heavy for our purposes both in terms of speed and complexity of user interface.

Perhaps the best option we plan to investigate in future research is a mixed approach: the interface would display only the text version of the sentence but it would internally know the (automatic) t-layer structure. Selecting any word that corresponds to the t-node of a verb would automatically extend the selection to all other belongings of the t-node, i.e. all auxiliaries of the verb. For role fillers, selecting any word from the role filler would select the whole t-layer subtree. In order to handle errors in the automatic t-layer annotation, the interface would certainly need to allow manual selection and deselection of words, providing valuable feedback to the automatic tools.

[3] http://ufal.mff.cuni.cz/pcedt2.0/
[4] http://ufal.mff.cuni.cz/treex/
[5] http://ufal.mff.cuni.cz/czeng/
[6] http://ufal.mff.cuni.cz/tred/

HMEANT	0.2833
METEOR	0.2167
WER	0.1708
CDER	0.1375
NIST	0.1167
TER	0.1167
PER	0.0208
BLEU	0.0125

Table 1: Kendall's τ for sentence-level correlation with human rankings.

3 An Experiment in English-Czech MT Evaluation

In this first study, we selected 50 sentences from the English-to-Czech WMT12 manual evaluation. The sentences were chosen to overlap with the standard WMT ranking procedure (see Section 3.1) as much as possible.

In total, 13 MT systems participated in this translation direction. We allocated 14 annotators (one annotator for the SRL of the reference) so that nobody saw the same sentence translated by more systems. The hypotheses were shuffled so every annotator got samples from all systems as well as the reference. Unfortunately, time constraints and the large number of MT systems prevented us from collecting overlapping annotations, so we cannot evaluate inter-annotator agreement.

Following Lo and Wu (2011a) and Callison-Burch et al. (2012), we report Kendall's τ rank correlation coefficients for sentence-level rankings as provided by a range of automatic metrics and our HMEANT. The gold standard are the manual WMT rankings. See Table 1.

We see that HMEANT achieves a better correlation than all the tested automatic metrics, although in absolute terms, the correlation is not very high. Lo and Wu (2011b) report τ for HMEANT of up to 0.49 and Lo and Wu (2011a) observe τ in the range 0.33 to 0.43. These figures are not comparable to our result for several reasons: we evaluated 13 and not just 3 MT systems, the gold standard for us are overall system rankings, not just adequacy judgments as for Lo and Wu (2011b), and we evaluate translation to Czech, not English. Callison-Burch et al. (2012) report τ for several automatic metrics on the whole WMT12 English-to-Czech dataset, the best of which correlates at $\tau = 0.18$. The only common metric is METEOR and it reaches 0.16 on the whole WMT12 set.[7] In line with our observation, Czech-to-English correlations reported by Callison-Burch et al. (2012) are higher: the best metric achieves 0.28 and averages 0.25 across four source languages.

The overall low sentence-level correlation of our HMEANT and WMT12 rankings is obviously caused to some extent by the problems we identified, see Section 4 below. On the other hand, it is quite possible that the WMT-style rankings taken as the gold standard are of a disputable quality themselves, see Section 3.1 or the detailed report on inter-annotator agreement and a long discussion on interpreting the rankings in Callison-Burch et al. (2012). Last but not least, it is likely that HMEANT and manual ranking simply measure different properties of MT outputs. The Kendall's τ is thus not an ultimate meta-evaluation metric for us.

3.1 WMT-Style Rankings

This section illustrates some issues with the WMT rankings when used for system-level evaluation. Obviously, at the sentence level, the rankings can behave differently but the system-level evaluation benefits from a large number of manual labels.

In the WMT-style rankings, humans are provided with no more than 5 system outputs for a given sentence at once. The task is to rank these 5 systems relatively to each other, ties allowed.

Following Bojar et al. (2011), we report three possible evaluation regimes (or "interpretations") of these 5-fold rankings to obtain system-level scores. The first step is shared: all *pairwise* comparisons implied by the 5-fold ranking are extracted. For each system, we then report the percentage of cases where the system won the pairwise comparison. Our default interpretation is to exclude all ties from the calculation, labelled "Ties Ignored", i.e. $\frac{\text{wins}}{\text{wins + losses}}$. The former WMT interpretation (up to 2011) was to include ties in both the numerator and the denominator, i.e. $\frac{\text{wins + ties}}{\text{wins+ties+losses}}$ denoted "$\geq$ Others". WMT summary paper also reports "$>$ Others" where the ties are included in the denominator only, thus giving credit to systems that are different.

As we see in Table 2, each of the interpretations leads to different rankings of the systems. More importantly, the underlying set of sentences also affects the result. For instance, the system ONLINEA jumps to the second position in "Ties Ignored" if we consider only the 50 sentences used in our HMEANT evaluation. To some extent, the differences are caused by the lower number of observations. While "All-No Ties" is based on 2893$\pm$134 pairwise comparisons per system, "50-No Ties" is based on just 186$\pm$30 observations. Moreover, not all systems came up among the 5 ranked systems for a given sentence. In our 50 sentences, only 7.3$\pm$2.1 systems were compared per sentence. On the full set of sentences, this figure drops to 5.9$\pm$1.7.

4 Problems of HMEANT Annotation

We asked our annotators to take notes and report any problems. On the positive side, some annotators familiar with the WMT ranking evaluation felt that in both phases of HMEANT, they "knew what they were doing and why". In the ranking task, it is unfortunately quite common that the annotator is asked to rank incomparably bad hypotheses. In such cases, the annotator probably tries to follow some subjective and unspoken criteria, which often leads to a lower in inter- and intra-annotator agreement.

On the negative side, we observed many problems of the current version of HMEANT, and we propose a remedy for all of them. We disregard minor technical issues of the annotation interface and focus on the design decisions. The only technical limitation worth mentioning was the inability to return to previous sentences. In some cases, this even caused the

[7] It is possible that Callison-Burch et al. (2012) use somewhat different METEOR settings apart from the different subset of the data.

Interpretation	Ties Ignored		≥ Others		> Others	
Sentences	All	50	All	50	All	50
cu-depfix	66.4	72.5	73.0	77.5	53.3	59.4
onlineB	63.0	61.4	70.5	69.3	50.3	49.0
uedin-wmt12	55.8	60.3	63.6	66.3	46.0	≀ 51.1
cu-tamch-boj	55.6	54.6	≀ 64.7	62.1	44.2	45.7
cu-bojar_2012	54.3	53.2	≀ 64.1	≀ 62.2	42.6	43.0
CU_TectoMT	53.1	≀ 54.9	60.5	59.8	≀ 44.6	≀ 49.0
onlineA	52.9	≀ 61.4	≀ 60.8	≀ 66.7	≀ 44.0	≀ 53.0
pctrans2010	47.7	≀ 54.1	55.1	≀ 60.1	40.9	≀ 47.1
commercial2	46.0	51.3	54.6	59.5	38.7	42.7
cu-poor-comb	44.1	41.6	≀ 54.7	50.5	35.7	35.2
uk-dan-moses	43.5	33.2	53.4	44.2	≀ 35.9	27.7
SFU	36.1	31.0	46.8	43.0	30.0	25.6
jhu-hiero	32.2	26.7	43.2	36.0	27.0	23.3

Table 2: WMT12 system-level ranking results in three different evaluation regimes evaluated either on all sentences or just the 50 sentences that were subject to our HMEANT annotation. The table is sorted along the first column and the symbol "≀" in other columns marks items out of sequence.

annotators to skip parts of the annotation altogether, because they clicked Next Sentence instead of the Next Frame button.

Note that the impact of the problems on the final HMEANT reliability varies. What causes just minor hesitations in the SRL phase can lead to complete annotation failures in the Alignment phase and vice versa. We list the problems in decreasing severity, based on our observations as well as the number of annotators who complained about the given issue.

4.1 Vague SRL Guidelines

The first group of problems is caused by the SRL guidelines being (deliberately) too succinct and developed primarily for English.

Complex predicates. Out of the many possible cases where predicates are described using several words, SRL guidelines mention just modal verbs and reserve a label for them (assuming that the main verb will be chosen as the Action, i.e. the predicate itself). This goes against the syntactic properties of Czech and other languages, where the modal verb is the one that conjugates and it is only complemented by the content verb in infinitive. Some annotators thus decided to mark such cases as a pair of nested frames.

The problem becomes more apparent for other classes of verbs, such as phasic verbs (e.g. "to begin"), which naturally lead to nested frames.

A specific problem for Czech mentioned by almost all annotators, was the copula verb "to be". Here, the meaning-bearing element is actually the adjective that follows (e.g. "to be glad to ..."). HMEANT forced the annotators to use e.g. the Experiencer slot for the non-verbal part of this complex predicate. In the negated form, "není (is not)", some annotators even marked the copula as Negation and the non-verbal part as the Action.

No verb at all. HMEANT does not permit to annotate frames with no predicate. There are however at least two frequent cases that deserve this option: (1) the whole sentence can be a nominal construction such as the title of a section, and (2) an MT system may erroneously omit the verb, while the remaining slot fillers are understandable and the whole meaning of the sentence can be also guessed. Giving no credit to such a sentence at all seems too strict. In some cases, it was possible for the annotators to find a substitute word for the Action role, e.g. a noun that should have been translated as the verb.

A related issue was caused by the uncertainty to what extent the frame annotation should go. There are many nouns derived from verbs that also bear valency. FGD acknowledges this and valency lexicons for Czech do include also many of such nouns. If the

Reference	Oblečky	musíme	vystříhat	z časopisů
Gloss	clothes	we-must	cut	from magazines
Roles	Experiencer	Modal	Action	Locative
Meaning	We must cut the clothes (assuming paper toys) from magazines			
Hypothesis	Musíme	vyříznout	oblečení z časopisů	
Gloss	We-must	cut	clothes from magazines	
Roles	Modal	Action	Experiencer	

Figure 2: An example of PP-attachment mismatch. While it is (almost) obvious from the word order of the reference that the preposition phrase "z časopisů" is a separate filler, it was marked as part of the Experiencer role in the hypothesis. In the alignment phase, there is no way to align the single Experiencer slot of the hypothesis onto the two slots (Experiencer, Locative) if the reference.

instructions are not clear in this respect, it is quite possible that one annotator creates frames for such nouns and the other does not, causing a mismatch in the Alignment phase.

PP-attachment. The problem of attaching prepositional phrases to verbs or to other noun phrases is well acknowledged in many languages including English and Czech. See an example in Figure 2.

A complete solution of the problem in the SRL phase will never be possible, because there are naturally ambiguous cases where each annotator can prefer a different reading. However, the Alignment phase should be somehow prepared for the inevitable mismatches.

Unclear role labels. Insufficient role labels. The set of role labels of HMEANT is very simple compared to the set of edge labels (called "functors") in the tectogrammatical annotation. Several annotators mentioned that the HMEANT roleset is hard to use especially for passive constructions or verbs with a secondary object.

Because the final HMEANT calculation requires aligned fillers to match in their role labels, the agreement on role labels is important. We suggest experimenting also with a variant of HMEANT that would disregard the labels altogether.

Other problematic cases are sentences where several role fillers appear to belong to the same type, e.g. Locative: "Byl převezen (He was transported) | do nemocnice (to the hospital) | v záchranném vrtulníku (in a helicopter)". While it is semantically obvious that the hospital is not in the helicopter, so this is not a PP-attachment problem, some annotators still mark both Locatives jointly as a single slot, causing the same slot mismatch. It is also possible that the annotator has actually assigned the Locative label twice but the annotation interface interpreted all the words as belonging to one filler only.

Coreference. The SRL guidelines are not specific on handling of slot fillers realized as pronouns (or even dropped pronouns). If we consider a sentence like "It is the man who wins", it is not clear which words should be marked as the Agent of the Action "wins". There are three candidates, all equally correct from the purely semantic point of view: "it", "the man" and "who".

A natural choice would be to select the closest word referring to the respective object, however, in constructions of complex verbs or in pro-drop languages the object may not be explicitly stated in the syntactically closest position. Depending on the annotators' decisions, this can lead to a mismatch in the number of slots in the subsequent Alignment phase.

Other problems. Some annotators mentioned a few other problems. One of them were paratactic constructions: the frame-labelling procedure does not allow to distinguish between sentences like "It is windy and it rains" vs. "It is windy but it rains", because neither "and" nor "but" are a slot filler. Similarly, expressions like "for example" do not seem to constitute a slot filler but still somehow refine the meaning of the sentence and should be preserved in the translation.

One annotator suggested that the importance of the SRL phase should be emphasized and the annotators should be pushed towards annotating as much as they can, e.g. also by highlighting all verbs in the sentence, in order to provide enough frames and fillers to align in the second phase.

Reference	Opilý řidič	těžce	zraněn
Gloss	A drunken driver	seriously	injured
Roles	Agent	Extent	Action
Meaning	A drunken driver is seriously injured.		
Hypothesis	Opilý řidič	vážně	zranil
Gloss	A drunken driver	seriously	injured (active form)
Roles	Agent	Extent	Action
Meaning	A drunken driver seriously injured (someone).		

Figure 3: A mismatch of the meanings of the predicates. Other roles in the frames match perfectly.

The following sections describe problems of the Alignment phase.

4.2 Correctness of the Predicate

HMEANT alignment phase allows the annotators to either align or not align a pair of frames. There is no option to indicate that the match of the predicates themselves is somewhat incorrect. Once the predicates are aligned, the user can only match individual fillers, possibly penalizing partial mismatches.

Figure 3 illustrates this issue on a real example from our data. Once the annotator decides to align the frames, there is no way to indicate that the meaning was reversed by the translation.

What native speakers of Czech also feel is that the MT output in Figure 3 is incomplete, an Experiencer is missing. A similar example from the data is the hypothesis "Svědek oznámil policii. (The witness informed/announced the police.)" The verb "oznámit (inform/announce)" in Czech requires the message (perhaps the Experiencer in the HMEANT terminology), similarly to the English "announce" but unlike "inform". The valency theory of FGD formally describes the problem as a missing slot filler and given a valency dictionary, such errors can be even identified automatically.

On the other hand, it should be noted that a mismatch in the predicate alone does not mean that the translation is incorrect. An example in our data was the phrase "dokud se současné umění nedočkalo ve Vídni nového stánku" vs. "než současné umění ve Vídni dostalo nový domov". Both versions mean "until contemporary art in Vienna was given a new home" but due to the different conjunction chosen ("dokud/než, till/until"), one of the verbs has to be negated.

4.3 Need for M:N Frame Alignment

The majority of our annotators complained that complex predicates such as phasal verbs or copula constructions as well as muddled MT output with no verb often render the frame matching impossible. If the reference and the hypothesis differ in the number of frames, then it is also almost certain that the role fillers observed in the two sentences will be distributed differently among the frames, prohibiting filler alignment.

A viable solution would be allow merging of frames during the Alignment phase, which is equivalent to allowing many-to-many alignment of frames. The sets of role fillers would be simply unioned, improving the chance for filler alignment.

4.4 Need for M:N Slot Alignment

Inherent ambiguities like PP-attachment or spurious differences in SRL prevent from 1-1 slot alignment rather frequently. A solution would be to allow many-to-many alignments of slot fillers.

4.5 Partial Adequacy vs. Partial Fluency

The original HMEANT Alignment guidelines say to mark an aligned slot pair as Correct or Partial match. (Mismatching slots should not be aligned at all.) A Partial match is described as:

> Role fillers in MT express part of the meaning of the aligned role fillers in the reference translation. Do NOT penalize extra meaning unless it belongs in other role fillers in the reference translation.

The second sentence of the instructions is probably aimed at cases where the MT expresses *more* than the reference does, which is possible because

the translator may have removed part of the content or because the source and the reference are both not quite literal translations from a third language. A clarifying example of this case in the instructions is highly desirable.

What our annotators noticed were cases where the translation was semantically adequate but contained e.g. an agreement mismatch or another grammar error. The instructions should exemplify, if this is to be treated as a Correct or Partial match. Optionally, the Partial match could be split into three separate cases: partially inadequate, partially disfluent, and partially inadequate and disfluent.

4.6 Summary of Suggested HMEANT Fixes

To summarize the observations above, our experience with HMEANT was overall positive, but we propose several changes in the design to improve the reliability of the annotations:

SRL Phase:

- The SRL guidelines should be kept as simple as they are, but more examples and especially examples of incorrect MT output should be provided.
- The Action should be allowed to consist of several words, including non-adjacent ones.
- The possibility of using automatic t-layer annotation tools should be explored, at least to pre-annotate which words form a multi-word predicate or role filler.

Alignment Phase:

- The annotator must be able to indicate a partial or incorrect match of the predicates themselves.
- Both frames as well as fillers should support M:N alignment to overcome a range of naturally appearing as well as spurious mismatches in the two SRL annotations.
- Examples of anaphoric expressions should be included in the guidelines, stressing that any element of the anaphora chain should be treated as an appropriate representant of the role filler.
- The Partial match could distinguish between an error in adequacy or fluency, or rather, the Alignment guidelines should explicitly provide examples of both types and ask the annotators to disregard the difference.

Technical Changes:

- The annotators need to be able to go back within each phase. (The division between the SRL and Alignment phases should be preserved.)

We do not expect any of the proposed changes to negatively impact annotation time. Actually, some speedup may be obtained from the suggested pre-annotation and also from a reduced hesitation of the annotators in the alignment phase thanks to the M:N alignment possibility.

5 Conclusion

We applied HMEANT, a technique for manual evaluation of MT quality based on predicate-argument structure, to a new language, Czech. The experiment confirmed that HMEANT is applicable in this setting, outperforming automatic metrics in sentence-level correlation with manual rankings.

During our annotation, we identified a range of problems in the current HMEANT design. We thus propose a few modifications to the technique and also suggest backing HMEANT with a linguistic theory of deep syntax, opening the avenue to automating the metric using available tools.

Acknowledgments

We would like to thank our annotators for all the comments and also Chi-kiu Lo, Karteek Addanki, Anand Karthik Tumuluru, and Avishek Kumar Thakur for administering the annotation interface. This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic) and the Czech Science Foundation grants P406/11/1499 and P406/10/P259 (Ondřej Bojar); and by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grant GRF621008 (Dekai Wu). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the RGC, EU, or DARPA.

References

Hervé Blanchon, Christian Boitet, and Laurent Besacier. 2004. Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals. In *Proceedings of International Conference on Spoken Language Processing ICSLP 2004*, Jeju Island, Korea, October.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar 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, July. Association for Computational Linguistics.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Charles J. Fillmore. 1968. The Case for Case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–90. New York.

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)*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2011a. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2011b. Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, pages 10–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Jarmila Panevová. 1980. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence]*. Academia, Prague, Czech Republic.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics. Submitted.

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.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Jana Šindlerová and Ondřej Bojar. 2010. Building a Bilingual ValLex Using Treebank Token Alignment: First Observations. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 304–309, Valletta, Malta, May. ELRA, European Language Resources Association.