A Fine-Grained Evaluation Method for Speech-to-Speech Machine Translation Using Concept Annotations

Robert S. Belvin HRL Laboratories 3011 Malibu Canyon Rd. Malibu, CA 90265 rsbelvin@hrl.com

Susanne Riehemann SRI International 333 Ravenswood Ave. Menlo Park, CA 94025 rieheman@ai.sri.com

Kristin Precoda SRI International 333 Ravenswood Ave. Menlo Park, CA 94025 precoda@speech.sri.com

Abstract

In this paper we report on a method of evaluating spoken language translation systems that builds upon a task-based evaluation method developed by CMU, but rather than relying on a predefined database of Interchange Format representations of spoken utterances, instead relies on a set of explicitly defined conventions for creating these interlingual representations. Our method also departs from CMU's in its scoring conventions in using a finer-grained approach to scoring (especially scoring of predicates). We have attempted to validate the legitimacy of this approach to speech-to-speech MT evaluation by looking for a relationship between the scores generated by this method, and the scores generated by a series of experiments using naïve human judgements of the meaning and quality of MT systems' output.

1. Introduction

In this paper we will describe the development of a method for evaluating a narrow domain speech-to-speech translation system and also discuss how the scores produced by that method relate to naïve human judgements about the quality of translations. The method is being developed as both a diagnostic tool and a performance metric for final system evaluation for the DARPA CAST program.[1] This program has as its goal the creation of two-way, speech-to-speech language translation systems for narrow domains, including "first encounter" medical care in field environments for Pashto, Farsi, Arabic, and Mandarin.

While there are certain widely accepted metrics for evaluating various components of a speech-to-speech MT system, there are considerably fewer widely accepted and well-understood metrics for evaluating end-to-end task-based speech-to-speech (S2S) MT, possibly none. The problem is particularly acute when comparing translations between different language pairs (as in CAST). Part of the problem is that end-to-end system evaluation depends so heavily on explicit user judgements of quality, which are hard to quantify (e.g. Nübel 1996).

It is understandable, then, why there should be such interest in methods and metrics for doing MT evaluation which are mostly or completely automated, *i.e.* methods in which no explicit user judgements are required. The BLEU/NIST method and utilities are currently the most important examples of this. But it isn't fully understood what all the factors are that influence a BLEU/NIST score. Several studies have shown that these metrics are very sensitive to the reference set used, and the relative rankings of two systems can be reversed under some circumstances (Culy & Riehemann 2003). To our knowledge, no one has studied how BLEU scores vary across languages.

We take the perspective that, although there are many aspects of the problem of carrying out effective S2S MT evaluation that are murky at best, there are also aspects that are *not* so mysterious, and we should exploit those to the best of our ability. For example, for a given source utterance and target translation, one can ask, and answer, questions such as:

i. is the translation in the target language an appropriate kind of speech/dialog act?
ii. is the core predicate for a particular source utterance clause appropriately expressed in the target language?
iii. are appropriate arguments associated with each core predicate?
iv. do the core predicate and its arguments comprise a coherent syntactic constituent?

Questions such as this are not terribly difficult to answer, although they do require some human effort and expertise. We do not mean to suggest that this is the *only* approach to carrying out S2S evaluation we should be pursuing, but it should be at least *one* of the approaches.

And, in fact, there are groups who have developed metrics along lines suggested in (i-iii) above, most notably CMU in their evaluation of JANUS, NESPOLE!, the C-STAR translator, and other S2S MT systems they and others have developed (Levin et al 2000). In particular, they have devised an evaluation method that essentially attempts to measure both accuracy and effectiveness simultaneously, by employing what they term a "practical interlingua" (not as rich but also not nearly so complex as most MT interlinguas—see Levin et al. 2000). The method evaluates the system in a task-oriented way, though with a finer-grained content accuracy metric than is found in most task-oriented evaluation methods. The interlingual structures used include both a speech-act label as well as *domain actions* (in-domain events and states) and the domain action's arguments.

The evaluation method used by CMU includes the following steps: the utterances from a machine translated bilingual dialogue are annotated with interlingual structures (which are referred to as Interchange Format (IF) structures); next a human tagger with understanding of both the IFs and the target language marks each *communicative goal* (each domain action and each argument are counted as distinct goals) as *succeed* or *fail*; finally, a score is assigned for each goal g_i in each

[1] The program was originally called "Babylon"

translated utterance (as in the formulae below), where *n* indicates the number of attempts made by the user to communicate the goal g_i:

$s(g_i) = 1/n$ (if the goal succeeds)

$s(g_i) = -(1-1/n)$ (if the goal fails)

The score for a given goal thus ranges between 1 and –1, with successful goal translations falling in the positive range and unsuccessful translations in the negative. Multiple attempts result in lower scores for both successful and failed translations.

2. Using Heuristics When There Is No Database

The method we have devised is based on this method, although we have departed from their method in several respects: (a) the CMU method appears to rely on the existence of a database of domain actions (essentially in-domain predicates and arguments) from which they could draw for tagging their utterances, but we do not anticipate the availability of such a database in our case; (b) we have refined the granularity of the scoring in some respects; and (c) we have simplified the structure of the interlingual structures themselves, in particular the structure of the arguments. As a way of achieving some degree of intercoder consistency, however, we have defined a set of detailed conventions that will offer a reasonable alternative to a pre-specified database of DAs. The first three conventions are shown below.

2.1 Some Conventions for Creating Interlingual Concept Representations

Convention 1: Arrangement of the arguments of verbal predicates

If there is more than one argument, then:

The first argument is the logical subject.[2]

The second argument is the logical object, or it is an oblique with the preposition explicitly marked (object of a preposition), or it is an embedded proposition.

The third argument is always a predicate of some kind, either an oblique or an embedded proposition, never a plain object.[3]

Example:

1) I hit the ball.
   hit(I,ball): Logical Subject=I and Logical Object=ball
2) I made Tom hit the ball
   cause(I,hit(Tom,ball)) Logical Subject = I,
   emb. prop = hit(Tom,ball)
3) I threw the ball to Tom
   PAST+throw(I,ball,to(Tom)) LogSubj = I,
   LogObj = ball, oblique = to(Tom)

Convention 2. Hyphens, underscores, and pluses

We have adopted the following conventions for symbols indicating links between elements:

dashes indicate a very tight connection (speech act names, multi-word expressions, verb plus particle),

underscores indicate a medium strength connection (compound nouns (e.g. *pain_medication*), certain adverbial phrases (e.g. *down_there)*),

pluses indicate the loosest (syntactic) connection (auxiliary plus main verb, quantifier plus noun, adjective plus noun; also the plus is generally used to concatenate items in the concept annotation which are not necessarily tightly connected syntactically (e.g. speech-act+main-predicate).

The primary difference between the three means of joining items is in how they are scored. This will be addressed in the later section on scoring conventions.

Convention 3. Reference-restricting elements

Reference-restricting elements on objects (like determiners, quantifiers, and adjectives) are concatenated with the object, e.g. this+language, six+weeks, your+right+leg, though we note also that in general we are recommending not indicating definiteness/indefiniteness on arguments (i.e. generally avoid using "the" or "a").

In all we have defined 14 of these conventions (thus far). A few examples of the concept representations created using these conventions are shown below:

Jalal were you given any pain medication by the paramedics?

request-info+give(paramedics,pain_medication,to(you))

Jalal are you allergic to any medication that you know of?

request-info+know(you,BE_allergic(you,
to(some+ medication)))

That's why you need the cast, to keep you from moving that part of your leg

give-info+need(you,cast)+PURPOSE+prevent(cast,you,
move(you,that+part_of(your+leg)))

We noted during our work on these conventions and in using them for evaluation purposes that one of the most challenging (though perhaps most important) goals one might achieve with them was the capturing of contextually supplied material in a way that other less labor-intensive methods would never be able to. For example, determining whether an appropriate translation was given for the following utterance taken in context could be very different than taken out of context:

all right left leg as I move down your thigh

The interlingual concept (or IL) representation we gave to this sentence was the following:

acknowledge+AND+request-info+INAL-HAVE(you,pain)
+in(left+leg)+WHILE+touch(i,your+thigh)+
consecutively+lower-down

The reason for this is clear if one looks at the discourse context for the utterance:

speaker 1: all right <UH> I'm gonna feel down your right leg

speaker 1: <UH> any pain here as I squeeze your leg as I move down <breath>

speaker 2: no right leg seems to be okay

[2] The notion of logical subject and logical object is a variant of the notion of "deep-subject" and "deep-object"; that is, surface perturbations may yield sentences in which the surface subject is not the logical or more agentive subject, such as in passives. In the annotations we are employing, the first argument of a multiple-argument predicate will always be the more agentive of the arguments.

[3] Obliques are treated simply as embedded predicates, e.g. the oblique argument in "Throw it to John" would simply be represented as *to(John)*, on analogy with the predicate in "John knows Bill" as *know(John,Bill)*.

speaker 1: <click> all right left leg as I move down your thigh

Note that with the degree of semantic decomposition we are employing it is relatively straightforward to indicate some part of the IL as contextually supplied, in case it is best to score such elements with a different weighting than explicitly stated elements. Exactly what the right way to indicate which part of the IL is inferred from prior context, and just how those elements of the IL should be scored, is an on-going research topic for us, which we will not attempt to address further here. However, we wish to note the potential for this approach to capture this kind of capability (or lack thereof) in a translation system, and we suggest contextually supplied information would be extremely problematic to capture with most simpler (*viz* knowledge-poor) approaches.

3. Scoring MT Output Using Interlingual Concept Representations

One way in which our approach also differs from the CMU method lies in how we score the translated utterances, which consists of giving partial, whole, or multiple point(s) for each element in the concept structure which is adequately expressed in the translation. For a given IL representation, we assign more elements a score than CMU's method, although the arguments of our IL representations tend to be simpler than theirs. The net difference in terms of granularity of scoring is unclear at this point. In addition to this difference, we weight them the concepts according to type of element: major predicates are assigned 2 points, simplex arguments are assigned 1 point, each part of a compound (such as "pain" in "pain medication") is assigned half a point, and so on. We have experimented with three different weightings thus far, and will continue to experiment with them and report on the results. Below are some examples of our scoring conventions:

3.1 Scoring Conventions[4]

Give 1 point for each item in the representation which is adequately represented in the target translation, with the following exceptions:

i. elements which are joined by an underscore each receive half a point if they are adequately represented in the target translation, i.e. "pain_medication" gets .5_5.
ii. elements which are joined by a dash receive only one point total if they are adequately represented in the target translation, i.e. "give-info" gets 1 point only.
iii. determiners/articles only get half a point. As mentioned, in general we are recommending not including (in)definite articles in the representations.
iv. each major predicate (essentially verbal predicates, or copula+predicate adjectives) should receive 2 points

To illustrate the use of some of these conventions, the following representation would be assigned points as indicated, as a representation of the input sentence "You will need to wear a cast for six weeks":

give-info+must+wear(you,cast)+for(six+weeks)
1 + 1 + 2(1,1) + 1(1+1)

A perfect translation of the input sentence should earn a score of 9. A machine-translated output sentence in the target language would be compared against this IL representation, and given a score between 0-9, which is then just converted into a percentage.

It is important to note that the + and () symbols in the scoring line are iconic with those in the IL representation, and should not be taken as mathematical symbols. That is, in order to make the scoring transparent for anyone who wishes to review the scores, as well as to prevent confusion for the scorers themselves during the scoring process, we preserve the various concatenation and bracketing symbols so there is a one-to-one correspondence between the elements in the IL and the elements in the scoring line. In order to arrive at a total, one simply adds all the numbers in the string (though with the provisos about elements joined with dashes and underscores described in (i-iii) just above). Thus, the above is tallied as 1+1+2+1+1+1+1+1=9. Two final points should be noted about the example, which are that the "for" predicate only receives one point, since it is only a modifier of the major predicate "wear", and the auxiliary "must" receives one point only for the same reason.

In addition to these conventions, we are recommending that the following questions be considered by the evaluators as they assign scores:

For a given predicate complex in the concept representation (predicate, arguments and modifiers),
i. is the predicate represented with appropriate words and grammatical structure?
if so give it full credit
ii. do grammar errors or word choice partially obscure the nature of the predicate or the relation of the arguments to each other, the event, state or relation expressed by the predicate?
if so give partial credit to the predicate, and full credit to each appropriately represented argument or modifier
iii. do grammar errors or word choice completely or mostly obscure the nature of the predicate or relations of the arguments to each other, the event, state or relation expressed by the predicate?
if so give no credit to the predicate, but give credit for each of the elements which are appropriately represented

4. Looking for Validation

Seeking external validation of our evaluation method as a final performance measure (*i.e.* not a diagnostic), we have begun collecting human judgements about the accuracy and effectiveness of a small set of machine translated utterances, and have explored the relation between these naïve human judgements and the scores assigned to the same utterances using our method. The procedure we used to collect these judgements consisted of extracting short but coherent fragments of in-domain (medical) dialogs (all in English) between role-playing doctors and (people acting as) patients, and then: (i) annotating the English sentences with the IL representations, (ii) running

[4] This list comprises the majority of the conventions, but is not complete due to length considerations.

the sentences through publicly available MT systems (web-accessible Systran and Power Translator), and (iii) collecting human judgements about the meaning and comprehensibility of each of the translated sentences (before allowing the raters to look at the original English sentences).

We collected human paraphrases and ratings of comprehensibility for 22 sentences, 10 produced by Systran and 12 from Power Translator. No rater saw the same sentence in the two translation versions. Raters were given different subsets of the 22 sentences. There were 25 raters, and a total of 210 ratings and paraphrases were collected. The paraphrases were used to see if the rater had actually understood the original sentence; we assigned 0, 0.5, or 1 ratings of "actual comprehension" according to whether we thought a paraphrase didn't reflect the meaning at all, or reflected partial or full comprehension, respectively.

A weakness of the data collection was that the description of the task given to the raters was defined in such a way that left too much leeway in interpretation, and was not completely consistent among the people administering the experiment. Thus, the human ratings of comprehensibility described below should not be taken to be the only ratings possible. This should be taken only as preliminary data which may not answer our questions beyond emphasizing that we need to be more rigorous and clever about how we conduct the experiments. The instructions given were:

> *Please translate the (machine translated) sentences below into "real English" using your best guess as to what they mean.*
> *Then rate the comprehensibility of the sentences, including how likely the sentence is to be misunderstood.*
> *Score the sentences on a scale from one to ten, with one being completely incomprehensible, and ten being perfectly comprehensible and accurate.*

We looked at perceived comprehension (i.e. the ratings that humans gave), the actual comprehension (as assigned by one of the experimenters), and two forms of concept annotation score, one which scored each element with either 0 or 1, and one which was a weighted concept annotation score (as described earlier). The two concept annotation scores were closely related, and neither shows much difference in its relation to other data. This is an area we want to explore further, as it may well be that different weighting schemes might yield very different results.

What we discovered was that, with the experimental method we employed, there was *no* relationship between the concept annotation scores and either the human ratings of comprehensibility, or the ternary ratings of whether a sentence was actually correctly understood or not. The human ratings of comprehensibility probably did not represent the information we really wanted, as the task was not well defined, and we believe it is likely that there is some way to gather human rater information that would show more of a relationship with the concept annotation scores. (That is, not finding a relationship here does not mean there is none to be found, just that these particular operationalizations of the concept annotation and the human ratings do not show one.)
The lack of relationship between the actual comprehension -- crudely indicated by how well the paraphrase captured the information in the original sentence -- and the concept annotation scores seems more interesting and more conclusive. However, again, concept annotation scores may contain useful information even if they do not relate to how liable to miscomprehension a sentence is. In particular, we expect that the concept annotation scores will be useful as a diagnostic for determining aspects of a translation system that are strong or weak.

5. Conclusion

In this paper we have reported on a method of evaluating spoken language translation systems. The method builds upon the task-based evaluation method developed by CMU, but, rather than relying on a predefined database of Interchange Format representations of spoken utterances, instead relies on a set of explicitly defined conventions for creating these interlingual representations. Inter-coder consistency does appear to be an issue, though more testing must be done in order to reach any firm conclusions on the extent to which reliability can be achieved. Our method also departs from CMU's in its scoring conventions, and in particular, we are using a finer-grained approach to scoring (especially scoring of predicates). Each small piece of the meaning of predicates in an utterance is scored, and the scoring of the elements is differentially weighted depending on the semantic contentfulness of the element being scored. We have attempted to validate the legitimacy of this approach to S2S MT evaluation by looking for a relationship between the scores generated by this method, and the scores generated by a series of experiments using naïve human judgements of the meaning and quality of MT systems' output. The results of the study (on the relation between our evaluation method and the evaluations given by naïve human judges) were, unfortunately, inconclusive, and pointed up the importance of careful experimental design in order to do this sort of validation.

6. Acknowledgements

We wish to thank Jeff Kuo, Beatrice Oshika, and the other members of the Babylon/CAST Evaluation Advisory Committee for helpful discussion in the development of this approach. This work was supported by the DARPA Babylon program, contract N66001-02-C-6023.

References

Culy, C. and S. Riehemann (2003) "The Limits of N-Gram Translation Evaluation Metrics," *Proceedings of the MT Summit IX*, AMTA.

Levin, L., Bartlog, B., Llitjos, A., Gates, D., Lavie, A., Wallace, D., Watanabe, T., and M. Woszczyna (2000) "Lesson Learned from a Task-Based Evaluation of Speech-to-Speech Machine Translation," *Proceedings of LREC 2000*.

Nübel, R. (1996) "End-to-End Evaluation in Verbmobil I," *Proceedings of MT Summit VI*.