Internal survey - qualitative analysis

Anonymized - February 2, 2022

Introduction

This document provides a first analysis of our own responses to the survey, which we ran internally prior to its distribution. At the end of the document is a response by one of the group members with some discussion points. We will contrast our own opinions with those of the respondents to the survey once the results are in.

Uses

The main use of error analysis is to understand system behaviour better than automatic metrics.
The direct benefit of error analysis is that we learn the **distribution of errors** (kind and frequency), which gives us a better sense of the overall **fidelity and accuracy** of the model. Going one step further, error analysis allows us to make more **meaningful comparisons** of the strengths and weaknesses of different systems (note: if the comparison between those systems is made using the same inputs). Once we have a concrete picture of the failure modes of different systems, it is also easier to **imagine the potential harms** these may cause. All of this enables us to get a better **sense of how/where systems could be improved**.

Barriers

There is no established taxonomy of errors, that NLG researchers and practitioners can use. This makes it harder to get started.
Developing annotation guidelines, the challenge is to obtain high enough inter-annotator agreement so as to have a reliable assessment of the errors made by the system. (Note: Though see recent work arguing in favour of having a multiplicity of different perspectives.)
Annotation in general is time-consuming. (Note: with standardisation of taxonomies and guidelines, we could collectively save a large amount of time.)

Missing

There is currently no standardisation of the error analysis process; particularly the lack of a standardised error taxonomy stands out. Tools might be useful, but it's possible to carry out an error analysis using existing tools as well.

Enabling factors
Most of us would really like to see stronger incentives to carry out an error analysis. For people who have never carried out an error analysis, it would be really useful to organise tutorials.

Requirements
The questionnaire already lists a number of requirements: error analyses should include a table with the distribution of errors in the output of their system. This data should be based on a formalised annotation procedure, with at least two annotators, so that the paper can also report inter-annotator agreement to gauge the reliability of the analysis. Beyond these requirements, our responses asked for:

- Details about the annotators (Note: see the paper on Data Statements for NLP.)
- Examples of system output, potentially tracing errors back to the source.
- Details on the exact methodology (how was the data selected, how was the taxonomy established, etc.)
- Definitions for each of the different error types, with full guidelines in the appendix.
- If applicable: statistics comparing different systems, e.g a Chi-square test to see if models significantly differ in the amount of specific kinds of errors.

Responses
Note that, besides this document, the responses may refer to the quantitative analysis of our own responses as well.

Response #1
I think that the arguments laid out in the Initial Survey document, on the uses of error analysis are good. They could perhaps be extended to note that error analysis is going to become inherently more useful as text length increases, due to questions on the overall text having less meaning because they are not tied to a span of text which is short enough for them to make sense. I am perhaps not explaining this well, but what I mean is that if we ask "Is the text grammatical", and the sentence is "I am a fish", there are only so many possible ways it can be wrong. If we look at three paragraphs of news, we have exploding combinations of candidate faults. If the texts are extremely short, error analysis and categorization of text are closer to being the same thing, although this is only a special case.

Having a standard taxonomy is nice in theory, but beware: https://

xkcd.com/927. It could be better to have some standard issues that are to be addressed by an annotation scheme, and ask researchers to detail what they did. This was covered in both this doc and the second one, there were a lot of interesting points on asking researchers to clearly define what they did, which in my opinion is more important than whether it is standardized. I think the variety of tasks under the NLG umbrella requires error analysis to have flexibility and simplicity in order for it to be adopted. I like the minimal definition of error analysis at the top of page 2/2 for this reason.

For the report on our answers. I do not think that the issue with tools is only in their existence, it is that there are deployment, maintenance, support, and annotator training costs (financial and/or research time). WebAnno was nice and configurable [...] but glitches/eccentricities in the interface, as well as the cost to us if anything went wrong with a hosted server that we had made available to crowd workers, was why we had workers annotate in MS Word then we transcribed to localhost WebAnno ourselves. A Javascript interface for within a crowdsource platform or a simple web server, but there is still the issue of development and the cost if it breaks. The old high tech vs low tech argument. Just some thoughts based on our experience.

I think we all seemed to agree on what should be done and the direction that the field should go, which is great, it is just some of the finer details that there may be some difference (standardized taxonomies vs case-specific following guidelines, etc). It will be good to see what those outwith our group think, and get an idea of what they might be willing to do. We are going to be quite biased.

Response #2
I think contrasting the barriers with the requirements will be interesting at the analysis phase. If people perceive barriers such as incentives to complete analyses but then have standards that in and of themselves act as a barrier (e.g. implying if I don't see this is in an error analysis when reviewing, I'd be less likely to consider it useful/accept), that may be a new barrier to consider.

I imagine we'll have many phrases that people use in free form responses that all map to the same general idea. We may need to decide later if we want to include any of the details of this in an Appendix so as to capture any potential nuances in intentioned meaning.