Analysis

March 2, 2022

```python
[1]: import pandas as pd
     from matplotlib import pyplot as plt
     import seaborn as sns
     import numpy as np

     import csv
     from collections import Counter
```

```python
[2]: def distribution(records, question):
         "Get distribution of answers, for a given question."
         c = Counter(record[question] for record in records)
         total = sum(c.values())
         empty = c['']
         counts = {key: {"number": value,
                         "percentage": (value/total) * 100,
                         "percentage_answered": (value/(total-empty)) * 100}
                   for key, value in c.items()}
         try:
             del counts[''] ['percentage_answered']
         except:
             pass
         return counts


     def get_questions(question, number):
         "Get questions for a range of questions in a grid."
         texts = []
         for i in range(1,number+1):
             item = f'Q{question}_{i}'
             text = QUESTIONS[item]
             text = text.split('-')[-1].strip()
             texts.append(text)
         return texts


     def get_texts(records, question):
         "Get answer texts."
```

```python
    texts = []
    for record in records:
        answer = record[question]
        identifier = record['ResponseId']
        if not answer == '':
            texts.append([identifier, answer])
    return texts


def basic_stats(records, question):
    "Print basic statistics about the results."
    counts = distribution(records, question)
    for key, results in counts.items():
        if not key == '':
            print(f"{key}: {results['number']} ({results['percentage_answered']:
 ↪.2f}%)")


def underscored(base, number, records):
    "Get answer distribution for all subquestions."
    results = dict()
    for i in range(1, number+1):
        question = f"Q{base}_{i}"
        results[question] = distribution(records, question)
    return results


def agreement(counts):
    "Select percentage answered for all answers except the empty string."
    results = dict()
    for answer in ['Strongly disagree', 'Somewhat disagree', 'Neither agree nor␣
 ↪disagree', 'Somewhat agree', 'Strongly agree']:
        try:
            results[answer] = counts[answer]['percentage_answered']
        except:
            results[answer] = 0
    return results


# No longer needed:
# def enumerate_ids(iterable):
#     "Enumerate iterable with zero-padded IDs."
#     for i, element in enumerate(iterable):
#         yield 'comment-' + str(i).zfill(3), element


def write_texts(texts, filename):
```

```
    "Write texts from a list to a file."
    with open('./texts/' + filename,'w') as f:
        writer = csv.writer(f)
        writer.writerow(['identifier', 'comment', 'code'])
        writer.writerows([row + ['Original comment'] for row in texts])
```

[3]:
```
df = pd.read_excel("[Distributed] Perceptions of Error Analysis_March 1,␣
 ↪2022_09.00.xlsx")
df = df.fillna('')
records = df.to_dict("records")
```

/Users███████pt/anaconda3/lib/python3.8/site-
packages/openpyxl/styles/stylesheet.py:221: UserWarning: Workbook contains no
default style, apply openpyxl's default
  warn("Workbook contains no default style, apply openpyxl's default")

[4]:
```
consented = [record for record in records if str(record['Q1 ']).
 ↪startswith("Yes")]

print(len(consented))

# For subgroup analysis:
academia = [record for record in records if str(record['Q2'])=='Academia']
industry = [record for record in records if str(record['Q2'])=='Industry']
```

67

[5]:
```
# If necessary, here are all questions:
QUESTIONS = records[0]
```

[6]:
```
QUESTIONS
```

[6]:
```
{'StartDate': 'Start Date',
 'EndDate': 'End Date',
 'Status': 'Response Type',
 'Progress': 'Progress',
 'Duration (in seconds)': 'Duration (in seconds)',
 'Finished': 'Finished',
 'RecordedDate': 'Recorded Date',
 'ResponseId': 'Response ID',
 'DistributionChannel': 'Distribution Channel',
 'UserLanguage': 'User Language',
 'Q1 ': 'Informed consent\n\n \n\nThis is the consent form for our study about
```
the status of error analysis in NLG. Full details about this study were provided
on the previous page. If you want to read this information again, you can go
back to the previous page. If anything is still unclear about this study, please
contact: C.W.J.vanMiltenburg@tilburguniversity.edu\n\n \n\nConsent\n\nBy

3

consenting, you indicate that you have read the description on the previous page, that you are voluntarily taking part in this study, and that you allow for your data to be processed. This means that:\n\n\n\tYou agree to your responses being anonymously recorded.\n\tYour answers will be used to study the status of error analysis in NLG, and may be used in future publications pertaining to this topic.\n\tThe data will be shared with our research team, with both local (hard drive) and online (protected cloud drive) backups. This data will be stored indefinitely, and made public upon completion of our research. Note again that none of your answers can be traced back to you.\n\tYou acknowledge that there is no financial compensation for taking part in this study.\n\n\n \n\nNote that you may still withdraw your consent after completing this form, without any negative consequences. We will delete all incomplete forms from our study.\n\n \n\nDo you consent?\n\nDo you agree to take part in this study? If you consent, please indicate this below by clicking "Yes". If you click "No", you will be directed to the end of this questionnaire. You may also close this page to stop participating in this study.',
 'Q2': 'Are you in academia or in industry? (If you have a dual affiliation, please respond with your dominant affiliation in mind.)',
 'Q3': 'How many years have you been working in NLG?',
 'Q4': 'Do you remember reading any NLG papers that include an error analysis?',
 'Q5': 'Did you find the error analyses to be useful?',
 'Q6': "What did you find useful about the error analyses you've seen?",
 'Q7': "Why didn't you find the error analyses to be useful?",
 'Q8': "Is it surprising to you that you haven't seen any published error analyses? - Selected Choice",
 'Q8_1_TEXT': "Is it surprising to you that you haven't seen any published error analyses? - Yes, because: - Text",
 'Q8_2_TEXT': "Is it surprising to you that you haven't seen any published error analyses? - No, because: - Text",
 'Q9': 'Have you ever carried out an error analysis?',
 'Q10': 'What did you find challenging or difficult about carrying out an error analysis?',
 'Q11': 'Did you feel like there were enough resources/reference material for you to carry out an error analysis?',
 'Q28': "Do you think you'll carry out an error analysis again in the future?",
 'Q29': 'Could you explain your answer to the previous question?',
 'Q12': 'Have you ever considered carrying out an error analysis?',
 'Q13': "What is the reason you haven't carried out an error analysis?",
 'Q14': 'Are you willing to carry out an error analysis?',
 'Q15': 'For what kinds of papers do you think error analyses may be useful?',
 'Q16_1': 'I would be more likely to carry out an analysis in a conference/journal paper if… - There was a higher page limit.',
 'Q16_2': 'I would be more likely to carry out an analysis in a conference/journal paper if… - There would be an existing error taxonomy that I could use.',
 'Q16_3': 'I would be more likely to carry out an analysis in a conference/journal paper if… - There would be dedicated annotation tools for

error analysis that I could use.',
 'Q16_4': 'I would be more likely to carry out an analysis in a conference/journal paper if… - There would be a crowdsourcing template for carrying out error analyses.',
 'Q16_5': 'I would be more likely to carry out an analysis in a conference/journal paper if… - Reviewers paid more attention to error analyses.',
 'Q16_6': 'I would be more likely to carry out an analysis in a conference/journal paper if… - There were an available pool of annotators or crowd workers',
 'Q16_7': 'I would be more likely to carry out an analysis in a conference/journal paper if… - I had more time.',
 'Q16_8': 'I would be more likely to carry out an analysis in a conference/journal paper if… - I had more money.',
 'Q16_9': 'I would be more likely to carry out an analysis in a conference/journal paper if… - I had more collaborators.',
 'Q17': 'Are there any other barriers that prevent you from carrying out an error analysis?',
 'Q18_1': 'Please indicate whether you agree or disagree with the following statements - There should be more error analyses in the NLG literature',
 'Q18_2': 'Please indicate whether you agree or disagree with the following statements - Error analyses are a valuable part of a paper.',
 'Q18_3': 'Please indicate whether you agree or disagree with the following statements - Carrying out an error analysis is enjoyable.',
 'Q18_4': 'Please indicate whether you agree or disagree with the following statements - Carrying out an error analysis is boring/tedious.',
 'Q18_5': 'Please indicate whether you agree or disagree with the following statements - Error analyses are necessary to fully evaluate the performance of an NLG system.',
 'Q18_6': 'Please indicate whether you agree or disagree with the following statements - Knowing what errors a system makes is helpful for future research.',
 'Q18_7': 'Please indicate whether you agree or disagree with the following statements - Knowing what errors a system makes is helpful for practitioners/NLG in industry.',
 'Q18_8': 'Please indicate whether you agree or disagree with the following statements - If you publish at a conference, and you present an NLG system as one of your main contributions, you should include an error analysis.',
 'Q18_9': 'Please indicate whether you agree or disagree with the following statements - If you publish in a journal, and you present an NLG system as one of your main contributions, you should include an error analysis.',
 'Q19': 'I am … likely to include an error analysis in a journal article than/as I would be for a conference publication.',
 'Q27': 'Please explain your answer to the previous question:',
 'Q20': 'Are there currently enough resources to support error analysis? - Selected Choice',
 'Q20_2_TEXT': 'Are there currently enough resources to support error analysis?

```
  - No, I am still missing: - Text',
  'Q21': 'Besides resources, are there any other factors that would make it more
likely for you to carry out an error analysis?',
  'Q23': 'What else would you recommend that authors should include in an error
analysis?',
  'Q24': 'This is the final question. Is there anything you would like to add or
comment on?'}
```

```
[7]: """
     TODO:
     - Subgroup analysis: academia vs industry
     - Heatmap tables
     """
```

[7]: '\nTODO:\n- Subgroup analysis: academia vs industry\n- Heatmap tables\n'

1 Demographics

```
[8]: # Where do people come from?
     basic_stats(consented, "Q2")
```

```
Academia: 51 (85.00%)
Industry: 8 (13.33%)
Other: 1 (1.67%)
```

```
[9]: # Time spent working in NLG:
     basic_stats(consented, "Q3")
```

```
6-10 years: 5 (8.47%)
Less than 2 years: 13 (22.03%)
2-5 years: 23 (38.98%)
11 or more years: 13 (22.03%)
I don't work in NLG: 5 (8.47%)
```

```
[10]: # Read an error analysis:
      basic_stats(consented, "Q4")
```

```
Yes: 33 (67.35%)
No: 16 (32.65%)
```

```
[11]: # Is it surprising that you haven't read an error analysis?
      basic_stats(consented, "Q8")
```

```
Yes, because:: 3 (42.86%)
No, because:: 4 (57.14%)
```

```
[12]: # Why is it surprising?:
      texts = get_texts(consented, 'Q8_1_TEXT')
      write_texts(texts, "surprising_because.csv")

      for ident, text in texts:
          print(ident, text)
          print('----')
```

R_2viGZF9YUjaTxAZ using NLG systems every day, I know that however good the
output is, it still makes small mistakes that need correcting such as confusing
"me" and "you" roles.

R_3PUctIwcmDJt5t4 Without an understanding of errors, especially regarding what
the most frequent errors involve, it is quite hard to correctly develop a
system. It may end up being just blind hyperparameter optimisation (for NN ones)

R_2cOi363Fh3IQN3w I haven't read plenty of literature as NLG is only part of my
research

```
[13]: # Why is it not surprising?:
      texts = get_texts(consented, 'Q8_2_TEXT')
      write_texts(texts, "not_surprising_because.csv")

      for ident, text in texts:
          print(ident, text)
          print('----')
```

R_3CClKJCQVTV5fKQ Page limit is often too less to report a detailed error
analysis. Also manually annotation of errors is very time consuming, even if the
training data is already manually annotated and the test data manually
evaluated.

R_2at4iBHFbZ9qXF9 it is time-consuming and tedious. Furthermore, it seems there
is disagreement about standards, so that results cannot be compared sensibly.

```
[14]: # Carried out an error analysis:
      basic_stats(consented, 'Q9')
```

Yes: 25 (67.57%)
No: 12 (32.43%)

```
[15]: # Willing to carry one out again (only people who answered 'yes'):
      basic_stats(consented, 'Q28')
```

Probably yes: 7 (30.43%)
Definitely yes: 15 (65.22%)

Probably not: 1 (4.35%)

```
[16]:  # Explanation for previous question
       texts = get_texts(consented, 'Q29')
       write_texts(texts, "carry_out_again_because.csv")

       for ident, text in texts:
           print(ident, text)
           print('----')
```

R_2uIYMUOPADUs0Gp Not on a formal and structured level as for now.

R_3EVTcFVWkcADmAn It can improve the results and quality of generations.

R_1Fb4cMbBLCr1Mps Need to measure accuracy of generated texts
----
R_2Bx0flTszYdxDPA They are too useful to not do them

R_tK6Arub4LFxhMTT Sometimes it is needed…

R_1OxiIuva6OIySjr I think it's useful

R_9FWWZPFa03RKtDb It is useful

R_8AMlQcW7cBOSOlH I think its important to manually inspect the data from a
computational linguistic perspective and it can provide valuable insights into
improving inputs perhaps or neural architectures in order to guarantee more
semantically adequate   production systems.

R_3PmNb56WDtzFlOO Essential part of evaluation!

R_1msf0KMkf86xvU7 I do believe in the importance of error analysis, so I'll make
sure to include them as much as possible. However, when working with
collaborators, not everyone sees its importance and is willing to invest time
(and resources) on it.

R_1oGS3sNlc9Pzbhu Scores from automatic evaluation metrics cannot reliably
detect or quantify all types of errors in NLG, so manual error analysis is still
probably the best kind of evaluation.

R_3kgFfd7vL38wXiE Why not :)

R_3dS2PTlM6FW2INE I feel like the experience with the error analysis I have
carried out helps me to outline better categories in the future.

R_2wmLQttiaGr9OkJ ?

R_1OrcJTKXKKERNbF I think carring out error analyses is a way to improve the

8

systems

R_u3vQ66FYA58WEnL Error analysis is crucial for many uses of NLG systems,
especially for systems intended to produce outputs for human audience.

R_12tjBK7tLWimKGc Not necessarily for published research, but doing so for my
team is a fairly regular part of my industry job

R_3rMofNwuc818Ldq I cannot imagine not having a look at the outputs to see what
went wrong.

R_12RQqpVGE2qYoq3 Alas, I'm too old to do so now, as other projects are more
important (writing a book, for example)
----

[17]:
```python
# Considered carrying one out (only people who answered 'no'):
basic_stats(consented, 'Q12')
```

Never: 4 (36.36%)
Once or twice: 2 (18.18%)
I'm planning to carry out an error analysis in the future: 4 (36.36%)
Regularly: 1 (9.09%)

[18]:
```python
# Willing to carry one out (only people who answered 'no'):
basic_stats(consented, 'Q14')
```

Probably yes: 3 (25.00%)
Definitely yes: 4 (33.33%)
Might or might not: 4 (33.33%)
Probably not: 1 (8.33%)

[19]:
```python
# Reasons for not doing it:
texts = get_texts(consented, 'Q13')
write_texts(texts, "reason_for_not_carrying_out.csv")

for ident, text in texts:
    print(ident, text)
    print('----')
```

R_2viGZF9YUjaTxAZ Whilst I do correct errors, I've never really considered
carrying out an error analysis.

R_3CClKJCQVTV5fKQ I didn't know of this option, I thought error description is
enough.

R_3PUctIwcmDJt5t4 I work in rule-based NLG, hence unexpected errors in output
are unlikely

R_1DAuvy72fx3Y9px My work was on errors in people not text

R_cIRoJzsDgiLpSWB My studies are taking much of my time

R_1Owpizm5kuD60On Hasnt been a part of my research problem, but is becoming
increasingly relevant

R_3Oprj30Ag4eTLZO The NLG systems I worked on were rule based, and before we
evaluated them we made sure that there were no errors. We carried out task-
based evaluations comparing different strategies, and sometimes asked users to
judge fluency etc but there were no generation errors to analyse.

R_2at4iBHFbZ9qXF9 see above (tedious, lack of standards)

R_2ZESZwXOU9XYLO6 I used to work in NLG and don't any more (sorry there was no
button for that at the beginning). I worked on rule-based system and there was
no error ananlysis of the NLG because it didn't make any errors, we made sure
the rules worked correctly before the NLG was used as part of a bigger system.
There may have been error analysis of e.g. the speech recognition errors which
led to the wrong response being generated, but not of the NLG itself.

R_2c0i363Fh3IQN3w I have not done a formal study for the NLG components of my
project so far.
----

2   Usefulness of error analyses

```
[20]: # Found useful:
      basic_stats(consented, 'Q5')
```

Moderately useful: 11 (35.48%)
Slightly useful: 2 (6.45%)
Very useful: 7 (22.58%)
Extremely useful: 11 (35.48%)

```
[21]: # What was useful about the analyses?:
      texts = get_texts(consented, 'Q6')
      write_texts(texts, "uses_of_error_analysis.csv")

      for ident, text in texts:
          print(ident, text)
          print('----')
```

R_2uIYMUOPADUsOGp General attention to the topic

R_1Fb4cMbBLCr1Mps general taxonomy gives idea of challenging aspects
----

R_1zRNqaef0umL2QF Qualitative analysis plus examples more informative than numeric scores for understanding where improvements are still needed
----
R_2Bx0flTszYdxDPA They help ground the limitations of the systems And create directions for future work
----
R_tK6Arub4LFxhMTT Allows to assess whether things you belive could be improved are indeed the things that should be
----
R_1OxiIuva6OIySjr They help understand the limitation of the proposed method.

R_9FWWZPFa03RKtDb It gives a complementary picture of standard metrics.

R_3CClKJCQVTV5fKQ Clear view on data and not only a cherry picking error description. The different categories are helpful to identify problems of the NLG system. NLG is often evaluated only by automatic metrics which are not perfect and do not take all issues into account, such a error analysis can help to identify errors and not only trying to reach the best scores. The NLG systems get more and more black boxes which we don't know what they are doing so a manual analysis is helpful for identifying problems which we had overlooked focusing on automatic metrics.

R_3PUctIwcmDJt5t4 Diversification of the errors, making the researcher able to fine-tune a system based on them.

R_8AM1QcW7cBOSO1H It provided useful insights into the type of errors produced in the output which are not evident in string based or corpus based metrics - BLEU, BLEURT etc
----
R_1DAuvy72fx3Y9px It shows the kinds of failures the system has, and points to where problems in the (often Black Box) System may lie.

It is also a good sign that the whole analysis of results has considered all results.

R_3PmNb56WDtzFlOO Qualitative insights about what doesnt work.   Which is useful for planning future research, and also for deciding whether to use a model/algorithm in a project (since some types of errors are much more concerning than others)

R_1msfOKMkf86xvU7 They provide an idea of where the models are actually failing. This helps understand if the proposed approach is tackling the problem it intends to solve (or improve on). It also gives ideas for future work.

R_1oGS3sNlc9Pzbhu Error analyses can show the areas where a system struggles to generate the correct output which can be accuracy, coherence, fluency, etc.

R_cIRoJzsDgiLpSWB The way that error is treated in language due to the

difficulty of the natural language processing methods and its variety across
the different levels of processes

R_1Owpizm5kuD60On Awareness of the kinds of biases these system exhibit when
trying to generate something meaningful and at the same point also showcasing
what needs to be done to further the improvement of these models
----

R_3dS2PTlM6FW2INE Providing extra depth about the state of their system. That
is, providing more detailed information about the strenghts and challenges
regarding the system that you would not be able to obtain if you would just
report average fluency, BLEU, etc. Furthermore, it is also more "objective" than
the average qualitative analysis that we often see, that generally just reports
overall first impressions of the author.

R_u3vQ66FYA58WEnL When NLG error analysis is conducted manually (i.e. by
humans), it provides valuable information about the quality of NLG, and also can
tell us where the problems are.

R_2c0i363Fh3IQN3w It specifies error providing categories for them.

R_3rMofNwuc818Ldq It clearly shows what are the system's flaws and strong sides,
and gives explicit hints with respect to how the system can be improved.

R_0fDpb4gKFzYr9PH understanding the weaknesses of the system

R_12RQqpVGE2qYoq3 I quite agree with   s conclusions in his paper:
                                   . Most error analyses I have seen are in
psycholinguistics (my background), and they allowed 'us' to develop an
architecture or pipeline, figuring out to some extent the time course of
language production. Actually, if I were allowed to take 3 books with me on an
island, then I'd choose Levelt's 'bible' Speaking, and one of the various books
devoted to errors, as they contain in hidden form how things work, and why they
don't. Errors are incredibly valuable clues, regardless the specific aspect of
the process (choice of meaning, word access, synctatic processing, …) or the
task (reading, writing, speaking, misunderstanding/mishearing/misreading).

```
[22]:  # For what kinds of papers are error analyses useful?:
       texts = get_texts(consented, 'Q15')
       write_texts(texts, "kinds_of_papers.csv")

       for ident, text in texts:
           print(ident, text)
           print('----')
```

R_2uIYMUOPADUsOGp KPI standards

R_3EVTcFVWkcADmAn Dialogue systems, response generations

R_1Fb4cMbBLCr1Mps all NLG research.

R_1zRNqaef0umL2QF most NLG papers

R_2Bx0flTszYdxDPA All papers with an experimental component

R_tK6Arub4LFxhMTT Ones proposing improvements, especially if the improvements
have a certain goal (e.g. changing syntax, did only the bottom line change? or
something syntactic)
Probably can think of other things

R_9FWWZPFa03RKtDb Any paper really, but I understand that it is not always
feasible, due to time and economic constraints. But it should be expected to a
higher extent than currently.

R_2viGZF9YUjaTxAZ I think an error analysis would be useful for any papers where
an NLG system is generating text, and making mistakes. I hadn't heard of error
analysis before this survey, but have read a lot of NLG papers, none of which
contained an error analysis. Not doing so, seems at best a little dishonest.

R_3CClKJCQVTV5fKQ All NLG paper with a small amount of test data so that the
manual error annotation is easy doable. For example, paper proposing a new NLG
system for machine translation, text simplification, text summarization,
question answering…

R_3PUctIwcmDJt5t4 E2E-NLG in particular, given the difficulties in debugging
black-box models.

R_8AMlQcW7cB0S0lH System papers - papers which claim to improve on semantic
adequacy/controllability for neural text generation systems i.e. pipeline neural
architecture and controllable neural generation, better input representation and
evaluation papers in general.

R_1DAuvy72fx3Y9px Any work on a system that outputs text

R_3PmNb56WDtzFlOO Any paper which evaluates a model, algorithm, or system should
include an error analysis

R_1msfOKMkf86xvU7 Pretty much all papers that claim to be doing something better
than others. For NLG, in particular, just showing that a model gets higher X
score(s) does not help understand why that is the case. It serves better to the
community to have an understanding of the real capabilities and limitations of
the models. So, to better compare systems, an error analysis can help show where
a particular system is "doing better" than the other, making a stronger case for
using it (or not).

R_1oGS3sNlc9Pzbhu - Papers describing novel approaches/architectures to NLG: It

would be useful to know whether a particular approach is prone to making
mistakes related to fluency, accuracy, coherency, etc.
- Papers comparing two or more NLG systems.

R_cIRoJzsDgiLpSWB Journal of Automated Reasoning
Journal of Intelligent Systems
Journal of Logic and Computation

R_3JhffzVm2xlXoiQ I think any NLG paper should conduct error analyses to help
readers better understand the limitations and potential risk of current model as
well as datasets.

R_1Owpizm5kuD6OOn Any paper which introduces a new NLG model should also talk
about the pitfalls of generations or the things that the model gets stuck in
since that provides an immediate reference to put things in perspective with.

R_3dS2PTlM6FW2INE Generally any paper that introduces a system in the NLG domain
(even more broadly: NLP). Especially if you have limited time and resources for
a quantitative human evaluation study, you can get interesting results with just
a small amount of annotators.

R_2wmLQttiaGr9OkJ explainable AI and confidence analysis

R_2ZESZwXOU9XYLO6 For papers using neural generation models which can produce
errors.

R_1OrcJTKXKKERNbF most of them: new models, new systems…

R_u3vQ66FYA58WEnL I have to say that what is considered an error depends
partially on the intended use of the NLG system. For systems that are intended
to produce grammatical coherent text for human audiences, error analysis is
necessary in order to get a good estimation of system quality. However, some
systems might be intended to produce different output, for example poetry, or
literary-style imitations, so criteria of errors may be different there. There
might also be papers that focus on computational efficiency, and thus disregard
quality of output, so those might avoid error analysis.

Some NLG papers use BLEU scores as indicators of NLG quality. This is very
convenient as BLEU scores can be computed automatically against available human-
sourced 'gold data'. But BLEU scores can be misleading and are not a good
alternative to manual analysis of errors.

R_2c0i363Fh3IQN3w All of them.

R_12tjBK7tLWimKGc Any paper that's presenting an NLG system, and many if not all
that survey multiple NLG papers or analyze evaluation methodologies (e.g. if
you're validating how well an automatic metric does, you might want to see how

it responds to different types of errors)

R_3rMofNwuc818Ldq Basically any paper that presents and/or compares outputs of (an) NLG system(s).

R_12RQqpVGE2qYoq3 Concerning papers, the problem lies mostly with the politics of the people in charge of these papers. I think that error analysis should definitely be taken on board, and we should try to spend more time reflecting what can be expected from them given a specific task.

```python
[23]:   # Reasons for disappointment:
        texts = get_texts(consented, 'Q7')
        write_texts(texts, "reasons_for_disappointment.csv")

        for ident, text in texts:
            print(ident, text)
            print('----')
```

3 Barriers and enabling factors

```python
[24]:   # Challenges:
        texts = get_texts(consented, 'Q10')
        write_texts(texts, "challenges.csv")

        for ident, text in texts:
            print(ident, text)
            print('----')
```

R_2uIYMUOPADUsOGp Scale and resources

R_3EVTcFVWkcADmAn It's time consuming and some times to cover all types of errors is very hard.

R_1Fb4cMbBLCr1Mps benchmarked against what? time-consuming. necessity for IRR (but usually lack of willing qualified participants)

R_1zRNqaef0umL2QF Time pressure

R_2Bx0flTszYdxDPA They can be time consuming to get right because it means contextualising your numbers
----

R_tK6Arub4LFxhMTT Defining categories,
Choosing amount of effort to invest and in what

R_10xiIuva6OIySjr It's not cool, so some of the co-authors had a push back
----

R_9FWWZPFaO3RKtDb It is hard to define clearly, especially in utput with poor quality, where the source of errors can be multiple. There is a lack of clearly described schemes, and the ones that exists are typically not well documented.

R_8AMlQcW7cBOSO1H The lack of clear methodology – type of errors.   Some appear random they pick a 100 and categorise the errors with not error schema. Often the sampling may or may not be statistically significant and there is no attempt to justify the sample size
----
R_3PmNb56WDtzF1OO Inter-annotator agreement. Ie, trying to define the error analysis well enough that different annotators produced comparable analyses.

R_1msfOKMkf86xvU7 Establishing a set of error categories that all annotators can understand and apply. This requires several iterations (just like with any annotation guideline). So, it is time consuming and tiring, specially if done only towards the end of the project.

R_1oGS3sNlc9Pzbhu It was time-consuming and prone to mistakes, especially when analysing for accuracy or correctness.

R_3kgFfd7vL38wXiE No previous experience in my area

R_3dS2PTlM6FW2INE There isn't really a standardized set of categories that you could use and build upon. So it felt like reinventing the wheel myself when trying to come up with a set of categories, going through the output.

R_2wmLQttiaGr9OkJ If not performed by human, an error analysis can require to process the generated output, such processing tools must be independent from the generator and be robust enough. Such tools does not exist for all languages. If processed by human -> usual hassle of time, recruitment and biais
----
R_1OrcJTKXKKERNbF - deciding the sample size
- defining the error categories: not too broad, not too fine grained
----
R_u3vQ66FYA58WEnL Manual error analysis is very time-consuming. In my case, we also used 2 annotators, for measuring inter-annotator agreement. I had to prepare the rubrics, then prepare sample materials for training annotators, then conduct training trials. Only then we could begin the real annotation and analysis.   So, again, it is a process that takes time and resources.

R_12tjBK7tLWimKGc It's not necessarily obvious what the categories used should be, especially when you have multiple similar/overlapping things going wrong.

R_3rMofNwuc818Ldq - Mainly, it can be tedious, since one needs to look at a lot of outputs.
- It can be challenging to define properly the error categories, which is crucial to make the analysis both informative (categories should be fine-grained enough) and reliable (categories should be result in a high IAA).

16

R_12RQqpVGE2qYoq3 I did this during my PhD (i.e., more than 40 years ago). We didn't have then the tools to do a decent job at the needed scale. Also, all the work was done manually.

```
[25]:  # Enough resources/reference materials at the time?
       basic_stats(consented,'Q11')
```

No: 13 (56.52%)
Yes: 10 (43.48%)

```
[37]:  answers = ['Strongly disagree', 'Somewhat disagree', 'Neither agree nor␣
       ↪disagree', 'Somewhat agree', 'Strongly agree']
       records = []
       for question, counts in underscored(16,9,consented).items():
           for answer in answers:
               percentage = 0
               if answer in counts:
                   percentage = counts[answer]['number'] # NOTE: Changed into number␣
       ↪rather than percentage!
               record = dict(question=question, answer=answer, percentage=percentage)
               records.append(record)

       df = pd.DataFrame(records)
       # Pivot to make a square table:
       df = df.pivot(index='question', columns='answer', values='percentage')
       # Reorder columns:
       df = df[['Strongly disagree', 'Somewhat disagree', 'Neither agree nor␣
       ↪disagree', 'Somewhat agree', 'Strongly agree']]

       plt.rcParams["figure.figsize"] = (15,4)
       ax = sns.heatmap(df,cmap=sns.light_palette("seagreen",␣
       ↪as_cmap=True),linewidth=1,cbar=False,annot=True)
       ax.xaxis.tick_top()
       plt.xticks(np.arange(5) + 0.5, labels=answers)
       plt.yticks(np.arange(9) + 0.5, labels=get_questions(16,9))
       plt.tick_params(top=False,left=False)
       plt.xlabel('')
       plt.ylabel('')
       plt.title("I would be more likely to carry out an analysis in a conference/
       ↪journal paper if…", y=1.2)
       plt.tight_layout()
       plt.savefig("Q16.pdf")
```

I would be more likely to carry out an analysis in a conference/journal paper if...

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
There was a higher page limit.	3	3	9	12	4
There would be an existing error taxonomy that I could use.	1	2	6	11	12
There would be dedicated annotation tools for error analysis that I could use.	1	4	7	10	10
There would be a crowdsourcing template for carrying out error analyses.	1	4	8	11	8
Reviewers paid more attention to error analyses.	0	2	6	9	15
There were an available pool of annotators or crowd workers	3	3	6	13	7
I had more time.	0	4	2	10	16
I had more money.	1	4	3	9	15
I had more collaborators.	0	3	7	10	12

```python
[27]:   # Other barriers?
        texts = get_texts(consented, 'Q17')
        write_texts(texts, "other_barriers.csv")

        for ident, text in texts:
            print(ident, text)
            print('----')
```

```
R_3EVTcFVWkcADmAn No
----
R_1Fb4cMbBLCr1Mps no
----
R_1zRNqaef0umL2QF Trade-off between improving model/approach vs conducting error
analysis
----
R_2Bx0flTszYdxDPA Time is the largest barrier
----
R_1OxiIuva6OIySjr Yes. It is not popular to conduct error analysis these days,
unfortunately.
----
R_2viGZF9YUjaTxAZ Not really.  I've just never considered running one.
----
R_3CClKJCQVTV5fKQ My supervisor do not see the relevance of it as we know no
paper with such a error analysis and hence they recommend me to not spend much
time on it.
----
R_8AMlQcW7cBOSOlH No
----
R_1oGS3sNlc9Pzbhu Reproducibility of error analysis or human evaluation in
general would be another concern, but there has been some interesting work
recently in this area.
----
R_1Owpizm5kuD6OOn The importance given to error analyses is very low in the
current research paradigm and we optimize for scores from reviewers for novelty
not for wrror analysis
----
```

R_3dS2PTlM6FW2INE Mostly time and resources, if there is a strict deadline, it is oftentimes quicker to just do a very straightforward quantitative analysis.

R_2wmLQttiaGr9OkJ lack of tools and resources for all languages
----
R_10rcJTKXKKERNbF I think error analysis should be carried out if possible by experts on the area and not by crowdworkers.

R_u3vQ66FYA58WEnL Time, money, availability of adequate annotators. A taxonomy of errors could be useful, but it cannot cover everything - there is a variety of issues that can be considered errors, including things that are errors only in some situation but not in other.. Also, a too-big of a taxonomy could be inconvenient to use.

R_3rMofNwuc818Ldq - My language level in the target language of the system I present.
- How many outputs are publicly available in other systems.

R_12RQqpVGE2qYoq3 nope
----

[28]: ```
# Enough resources/reference materials currently?
basic_stats(consented,'Q20')
```

No, I am still missing:: 20 (66.67%)
Yes: 10 (33.33%)

[29]: ```
# What is still missing?
texts = get_texts(consented, 'Q20_2_TEXT')
write_texts(texts, "missing.csv")

for ident, text in texts:
    print(ident, text)
    print('----')
```

R_2uIYMUOPADUsOGp Standards

R_1Fb4cMbBLCr1Mps the ability to outsource error analysis!

R_9FWWZPFaO3RKtDb Better documented taxonomies and procedures
----
R_2viGZF9YUjaTxAZ Funding. Whilst an error analysis is important, the way I work would mean that performing an error analysis would take away time from working on the NLG system itself, which could in turn reduce errors made. Funding could help this.

R_3CClKJCQVTV5fKQ Knowledge on the topic. I don't which resources exist yet.

R_3PUctIwcmDJt5t4 Don't know which tool is missing, but the practice itself
seems relatively novel, so I can expect new resources coming in the future.

R_1msf0KMkf86xvU7 Error analysis taxonomy, best practices, guidance, annotation
tools.

R_3JhffzVm2xlXoiQ An efficient guideline and platform for setting the standard
and replicable error analysis.

R_2at4iBHFbZ9qXF9 taxonomies, examples how to use them, tools
----
R_3dS2PTlM6FW2INE A good taxonomy of error categories you could typically use.

R_u3vQ66FYA58WEnL I am not aware of widely recognized resources for error
analysis in NLG.

R_2c0i363Fh3IQN3w Funding to do a study involving native speakers (I am not a
native English speaker)

```
[30]:  # Other factors that make it more likely for you to carry out an error analysis?
       texts = get_texts(consented, 'Q21')
       write_texts(texts, "enabling.csv")

       for ident, text in texts:
           print(ident, text)
           print('----')
```

R_2uIYMUOPADUsOGp Automation

R_3EVTcFVWkcADmAn No

R_1Fb4cMbBLCr1Mps perhaps, but I cannot think of one at the moment
----
R_1zRNqaef0umL2QF More explicit recognition of value of error analysis in review
forms

R_1OxiIuva6OIySjr Having it as a requirement. Or ideally making it more common,
as a practice that is considered valuable, and then researchers would be
inclined to do it not bc of a requirement.

R_9FWWZPFa03RKtDb More money for research, and easier to higher short-term
staff.

R_2viGZF9YUjaTxAZ No, on balance I think it's something that should be done when
presenting a paper or a conference talk.

R_3CClKJCQVTV5fKQ More acceptance of error analysis in the NLP community.

R_8AMlQcW7cBOSOlH Time and resources (to support the activity)

R_1DAuvy72fx3Y9px This is mostly a problem with research culture, when this (error analyses) becomes normal resources will be produced to fulfil the need for resources.

R_1msfOKMkf86xvU7 More money

R_1Owpizm5kuD6OOn A thorough task specific taxonomy which easily helps immediately attribute errors found in a certain bucket and also helpful since anything not belonging to the buckets is important and worthy to look at

----

R_2at4iBHFbZ9qXF9 money dedicated for persons doing the analyses, and also money for instructing them clearly in order to get comparable results

----

R_3dS2PTlM6FW2INE Having more experience with carrying it out would limit the amount of time necessary to set one up.

R_u3vQ66FYA58WEnL The major factor is that researchers must be acutely aware of the importance of error analysis for research, and for industrial/market applications..

R_2cOi363Fh3IQN3w Certainly, if my paper, concentrating on a very small, yet precise fragment, would have the same probability of acceptance of a paper that uses ML on a broad, yet very unprecise linguistic boundaries.

R_12RQqpVGE2qYoq3 In my case, no, because, as mentioned, given my age (nearly 74), I have other priorities.

----

4 General opinions

```
[31]: def insert(word, phrase, position):
          "Insert token into phrase, at position."
          tokens = phrase.split()
          new_tokens = tokens[:position] + [word] + tokens[position:]
          new_phrase = ' '.join(new_tokens).replace(' \n ','\n')
          return new_phrase


      modified_questions = get_questions(18,9)
      modified_questions[4] = insert('\n', modified_questions[4],7)
      modified_questions[5] = insert('\n', modified_questions[5],8)
      modified_questions[6] = insert('\n', modified_questions[6],8)
      modified_questions[7] = insert('\n', modified_questions[7],13)
      modified_questions[8] = insert('\n', modified_questions[8],13)
      modified_questions
```

```
[31]: ['There should be more error analyses in the NLG literature',
       'Error analyses are a valuable part of a paper.',
       'Carrying out an error analysis is enjoyable.',
       'Carrying out an error analysis is boring/tedious.',
       'Error analyses are necessary to fully evaluate\nthe performance of an NLG
      system.',
       'Knowing what errors a system makes is helpful\nfor future research.',
       'Knowing what errors a system makes is helpful\nfor practitioners/NLG in
      industry.',
       'If you publish at a conference, and you present an NLG system as\none of your
      main contributions, you should include an error analysis.',
       'If you publish in a journal, and you present an NLG system as\none of your
      main contributions, you should include an error analysis.']
```

```
[39]: answers = ['Strongly disagree', 'Somewhat disagree', 'Neither agree nor␣
      ↪disagree', 'Somewhat agree', 'Strongly agree']
      records = []
      for question, counts in underscored(18,9,consented).items():
          for answer in answers:
              percentage = 0
              if answer in counts:
                  percentage = counts[answer]['number']
              record = dict(question=question, answer=answer, percentage=percentage)
              records.append(record)

      df = pd.DataFrame(records)
      # Pivot to make a square table:
      df = df.pivot(index='question', columns='answer', values='percentage')
      # Reorder columns:
      df = df[['Strongly disagree', 'Somewhat disagree', 'Neither agree nor␣
      ↪disagree', 'Somewhat agree', 'Strongly agree']]
```

```python
plt.rcParams["figure.figsize"] = (15,4)
ax = sns.heatmap(df,cmap=sns.light_palette("seagreen",␣
 ↪as_cmap=True),linewidth=1,cbar=False,annot=True)
ax.xaxis.tick_top()
plt.xticks(np.arange(5) + 0.5, labels=answers)
plt.yticks(np.arange(9) + 0.5, labels=modified_questions)
plt.tick_params(top=False,left=False)
plt.xlabel('')
plt.ylabel('')
#plt.title("...", y=1.2)
plt.tight_layout()
plt.savefig("Q18.pdf")
```

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
There should be more error analyses in the NLG literature	0	1	1	10	19
Error analyses are a valuable part of a paper.	0	0	2	4	25
Carrying out an error analysis is enjoyable.	0	7	6	14	3
Carrying out an error analysis is boring/tedious.	3	4	6	17	0
Error analyses are necessary to fully evaluate the performance of an NLG system.	1	0	1	5	23
Knowing what errors a system makes is helpful for future research.	0	0	0	9	21
Knowing what errors a system makes is helpful for practitioners/NLG in industry.	0	0	1	5	24
If you publish at a conference, and you present an NLG system as one of your main contributions, you should include an error analysis.	0	0	5	13	12
If you publish in a journal, and you present an NLG system as one of your main contributions, you should include an error analysis.	0	0	2	10	18

```python
[33]: # More/less/equally likely to include error analysis in journal
      basic_stats(consented, 'Q19')
```

Equally: 14 (48.28%)
More: 15 (51.72%)

```python
[34]: # Explanation for previous question:
      texts = get_texts(consented, 'Q27')
      write_texts(texts, "explanation_journal_preference.csv")

      for ident, text in texts:
          print(ident, text)
          print('----')
```

R_3EVTcFVWkcADmAn It's make my NLG system more accurate.

R_1Fb4cMbBLCr1Mps Based on assumption that the conference is NOT ACL or similiar
top-tier venue!

R_1zRNqaef0umL2QF More space, less deadline pressure

R_2Bx0flTszYdxDPA The difference would be the depth of the analysis but heading
something remains Important in both cases

R_tK6Arub4LFxhMTT The quality of my work should not differ by the venue

----

R_1OxiIuva6OIySjr I think it should be a norm, similar to reporducibility, at both venues.

R_9FWWZPFaO3RKtDb There are more space and higher expectations.

R_2viGZF9YUjaTxAZ Having discovered that error analysis is a thing, if I were in the position of writing a paper or presenting at a conference, I would almost certainly include an error analysis.

R_3CC1KJCQVTV5fKQ Higher page limit, more time to publish, more quality and higher demands in journal articles

----

R_3PUctIwcmDJt5t4 Journal article usually offer a higher page limit.

R_8AMlQcW7cBOSOlH I think for long papers it is essential and I would expect this any ACL endorsed conference.
Also I think there is an overreliance on metric based evaluations without any clear understanding of there strengths and weakness(BLEU) and a general decline to provide any linguistic analysis of outputs. I fear the focus on deep learning approaches has resulted in a decline of computational linguistic skills in postgraduate researchers (with solely an ML training) and nor are such research evaluation encouraged. In some cases they don't have the linguistic skills.

R_1DAuvy72fx3Y9px More space and time involved in a journal publication, means it is more likely to have an error analysis.

R_3PmNb56WDtzFlOO Error analyses should be included in both conf and journal papers

----

R_1msfOKMkf86xvU7 Normally, having more space in a journal article allows to expand on the experiments, including the error analysis. However, since our research are is heavily conference-focused, our main points of reference are conference papers. So, since they are regarded in a higher standard, they should also be subjected to a higher level of scrutiny and quality expectation. So, I do think an error analysis should be included in both.

R_1oGS3sNlc9Pzbhu The deadlines and page limits for conference papers are typically tighter, so I would be slightly less likely to do this a conference paper.

R_cIRoJzsDgiLpSWB Error analysis will contribute to may article definitely because it can increase the what my investigation has done.

R_2at4iBHFbZ9qXF9 Journal articles are more comprehensive, and they should include an error analysis to give a more complete picture of an NLG system.

R_3dS2PTlM6FW2INE Bigger page limit, you are not as much faced with a strict
deadline that requires you to think about how to do the most in the least amount
of time.

R_2wmLQttiaGr9OkJ depend on size and the study. If the system is for
legal/medical purpose or for creativity

R_u3vQ66FYA58WEnL journal reviewers tend to be more demanding on details,
including error analysis. Also, journals provide more space (pages) for
articles.

R_2c0i363Fh3IQN3w I understand in our field conference paper are as important as
journal papers.

R_3rMofNwuc818Ldq I understand that there is generally more space in a journal
paper, but on needs to find space in a shorter paper too.

R_12RQqpVGE2qYoq3 Difference of space and time.
----

5   Requirements for reports of error analyses

```
[35]:   texts = get_texts(consented, 'Q23')
        write_texts(texts, "reporting_requirements.csv")

        for ident, text in texts:
            print(ident, text)
            print('----')
```

R_3EVTcFVWkcADmAn   May be a table that show correlation between different types
of errors.

R_1Fb4cMbBLCr1Mps sufficient evidence to evaluate the standard of the human
rater(s)

R_1zRNqaef0umL2QF Better to use a sensible characterization of errors that
actually occur rather than trying to shoehorn them into an existing taxonomy
----
R_1OxiIuva6OIySjr This is important. Thanks!

R_2viGZF9YUjaTxAZ I think including the types of errors made is fine, however, I
think that a list of the errors would be incredibly useful, as this would allow
people reading papers to see commonality amongst NLG systems (for example, do
GPT-J, GPT-3 and GPT-NeoX all make the same mistakes?)

R_3CClKJCQVTV5fKQ Annotation schema in which they explain their error categories

if adapted from another NLG task, e.g., question answering might require other categories than machine translation.
Not only raw numbers, maybe percentages would be better. Also naming which categories were ignored and due to which reasons.

R_8AM1QcW7cBOSOlH statistically driven sampling (stratified where appropriate even)

R_1DAuvy72fx3Y9px If (real) users find the system helps in a (real) task.

R_3PmNb56WDtzFlOO Annnotation process should be described in enough detail that other researchers can replicate the analysis and get similar results

----

R_1msfOKMkf86xvU7 The annotation guidelines, and the process followed to train the annotators. This can help with adopting a similar methodology for papers on the same task that aim to compare against them.

R_1oGS3sNlc9Pzbhu Types of errors and how that impacts a system. E.g., a system which generates fluent and grammatically correct output but contains factual error is not very useful.

R_3dS2PTlM6FW2INE A description of how the authors created the categories, with some opportunity for the annotators to report their satisfaction with the aplicability of the categories.

R_u3vQ66FYA58WEnL Proper metrics for measuring inter-annotator agreement. This is an issue not only in NLG. There is a variety of metrics and they are not all well-known or properly used.

However, I also warn aginst over-formalising error analysis!

R_2c0i363Fh3IQN3w You are thinking on a very large output. Sometimes you don't have that data. Have you considered the necessary cardinality of your sample to have a trustable agreement coefficient? The characteristics of your raters are also important, if you do this, though, the composition of your annotator group and the quantity of your sample shall vary. All this is quite costly.

R_3rMofNwuc818Ldq Ideally some comparison with errors in concurrent systems if not already available.

R_12RQqpVGE2qYoq3 some of the hypotheses or ideas that came to their mind when coming across the error/mistake

----

6 General comments

```
[36]: texts = get_texts(consented, 'Q24')
      write_texts(texts, "general_comments.csv")

      for ident, text in texts:
          print(ident, text)
          print('----')
```

R_2uIYMUOPADUsOGp Thanks

R_3EVTcFVWkcADmAn No

R_1Fb4cMbBLCr1Mps no

R_1zRNqaef0umL2QF Talk to PCs about review forms

R_2viGZF9YUjaTxAZ No. Thank you for introducing me to the concept of error
analysis. In at least some small way, I will probably take this concept and use
it in our work.

R_3CClKJCQVTV5fKQ I didn't know much on error analysis before answering the
questionnaire, hence, I couldn't rate the amountzof existing error analysis
tools. I would have liked to have a "I don't know" field for the Likert scale
questions.

R_8AMlQcW7cBOSOlH This is an important study

R_1msfOKMkf86xvU7 Thanks for carrying out this survey. Looking forward to the
results and the recommendations.

R_2wmLQttiaGr9OkJ thanks for doing such study

R_10rcJTKXKKERNbF Error analysis should focus on language features, text genre
characteristics and adequacy to the task, not a mere statistical analysis.

R_u3vQ66FYA58WEnL Thank you, it was a quite good survey.

R_2c0i363Fh3IQN3w I think the study of NLG errors should also take into account
the structure of the sentence and it is important the study is human based.

R_3rMofNwuc818Ldq Looking forward to the paper on the results of the study!

R_12RQqpVGE2qYoq3 It seems to me that there is a fundamental difference
concerning error analysis depending on whether we look at online processing or
written discourse. This is sth that I didn't see mentioned at all in this
questionaire, which otherwise is quite good.

[ ]: