Rethinking the Agreement in Human Evaluation Tasks (Position Paper)

Jacopo Amidei and **Paul Piwek** and **Alistair Willis**
School of Computing and Communications
The Open University
Milton Keynes, UK

Abstract

Human evaluations are broadly thought to be more valuable the higher the inter-annotator agreement. In this paper we examine this idea. We will describe our experiments and analysis within the area of Automatic Question Generation. Our experiments show how annotators diverge in language annotation tasks due to a range of ineliminable factors. For this reason, we believe that annotation schemes for natural language generation tasks that are aimed at evaluating language quality need to be treated with great care. In particular, an unchecked focus on reduction of disagreement among annotators runs the danger of creating generation goals that reward output that is more distant from, rather than closer to, natural human-like language. We conclude the paper by suggesting a new approach to the use of the agreement metrics in natural language generation evaluation tasks.

1 Introduction

Human evaluations are broadly thought to be more valuable the higher the inter-annotator agreement (IAA). In natural language processing (NLP), IAA is often viewed as a means of assessing the quality of data on a task, in particular, the reliability. Also, it helps to identify poor annotations, provide information on where annotation guidelines can be improved, and establish an upper bound on system performance. Nevertheless some studies, for example (Sampson and Babarczy, 2008), (Kovář 2016) and (Kovář et al., 2016)[1], propose IAA limitations when used in the annotation of natural language generation (NLG)[2], where annotation usually concerns the quality of a generated sentence – including a range of features such as syntax, semantic and pragmatics. Indeed, natural language brings in itself a variability which cannot be reduced, except by weakening its expressive power. For instance, the perception of sentence idiomaticity can change from person to person because of styles, educational and regional difference (for example differences between British and American English for English sentences). Far from being a problem this shows the richness, variability and beauty of human languages.

In this paper we are going to study IAA in the area of Automatic Question Generation (AQG)[3]. We performed experiments which show that there are factors inherent to question annotation that place an upper-bound on IAA. We found subjective bias that cannot be removed by improving the annotation

[1] It is also worth mentioning Martinez Alonso et al. (2016), where the authors use the systematic disagreement in order to train their model. Martinez Alonso et al. raise the discussion about how to distinguish and use potentially relevant from irrelevant disagreement. Although this problem is very interesting and deserves a deep analysis, it goes beyond the scope of the present paper.

[2] In what follows, when we refer to NLG tasks, we consider tasks that require generation of human language and whose possible output space is, as a result, very wide (given that a natural language often allows the same information to be expressed in many different ways). Input may vary, and range from text and data to images. Examples of these are: dialogue generation, question generation, image caption generation, narrative generation, and so on.

[3] Although we focused on the AQG task, we believe the causes we discovered behind annotators' disagreement are more general, and they are likely to apply to different NLG tasks.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3318–3329
Santa Fe, New Mexico, USA, August 20-26, 2018.

guidelines, unless we completely distort the goal of producing human language variability[4]. Our empirical study provides new evidence in strong support of the claim that high IAA is not always desirable. This work shows that for AQG tasks, a low IAA in the evaluation phase is part of the nature of language. More generally, the experimental results suggest that in NLG the ability to automatically generate sentences as similar as possible to human generated ones seems to go in the opposite direction from the achievement of a high IAA evaluation score. Although the Kappa coefficients agreement, normally used for measuring IAA, takes into account some level of randomness introduced by annotators, it might be that the Kappa scales of interpretation used to check IAA are not generally adequate for NLG purposes. Accordingly, we propose to rethink the agreement in human evaluation tasks.

2 Related work and preliminary concepts

Since Carletta's (1996) paper, which introduced the use of the kappa statistic in NLP, several papers have studied the IAA in NLP tasks. Among those, we now describe the most relevant to our work.

Bayerl and Paul (2011) investigate different factors which impact the IAA score, performing a meta-analytic investigation that involves 96 annotation studies. From their analysis, the authors conclude that at least seven factors are determinant in affecting the IAA values. These factors are: "annotation domain, number of categories in a coding scheme, number of annotators in a project, whether annotators received training, the intensity of annotator training, the annotation purpose, and the method used for the calculation of percentage agreements". They conclude their study giving some suggestions to decrease annotator disagreement. Ng et al. (1999) study IAA in the field of word sense disambiguation. The authors found that we can increase the IAA by collapsing sense classes. Indeed, they report that, for the 53 nouns they examined, reducing the nouns' senses from an average of 7.6 per noun to an average of 4, raises the average Kappa score (computed as an average of the Kappa value of the individual nouns) from 0.463 to 0.862. In order to improve the evaluation IAA of the task B defined in the Shared Evaluation Task Challenge (QG -STEC) (Rus et al., 2012), Godwin and Piwek (2016) define an interactive process where the annotators can discuss their opinions about the criteria used in the evaluation. At the end of the evaluation process, repeated three times with three annotators, they achieved high IAA with a pick of 0.94 for one of the five criteria used in the evaluation. In recent work, Novikova et al. (2018), use a rank-based magnitude estimation methodology in order to improve the agreement of human judgements.

The papers we have just presented share the assumption that a high IAA is a desirable goal and they study some ways to increase it. In this paper we reject the assumption that arbitrarily high IAA is always desirable or obtainable. Our paper is more in line with the direction of Sampson and Babarczy (2008), who investigate the upper bounds that can be achieved on IAA. More precisely "the limits to the potential precision of English grammar annotation" are studied. Sampson and Babarczy perform their experiment using the SUSANNE scheme (Sampson, 1995) for a parse-tree structure task. From their study they conclude that discrepancy in IAA emerges for three main reasons: A) violation of an explicit feature of the annotation scheme; B) the lack of a single, unambiguous annotation decision yielded by the scheme, even though the meaning of the text is clear; C) structural ambiguity in the text.

A really interesting survey for understanding of the IAA in NLP is presented by Artstein and Poesio (2008), who give an excellent analysis of the K agreement measures. The authors discuss the mathematics and interpretation of these coefficients and their use in several computational linguistic tasks.

Automatic Question Generation task definition: Automatic question generation is the task "*of automatically generating questions from various inputs such as raw text, database, or semantic representation*" (Rus et al., 2008). This definition, adopted by the AQG community, leaves room for researchers to decide which kind of questions and which kind of input to work with. Following Piwek and Boyer (2012), we can characterize a specific AQG task as relying on three features: what the input is, what the output is and finally the relation between the input and the output.

[4]Since human evaluations are valuable tools for generative system developers — indeed the evaluation results can be used for error analysis — artificial restrictions on annotation schemes can bias system developers towards ignoring important aspect of human language. This problem can be more accentuated if the annotation guidelines are used in order to create a corpus for training purposes.

Given the wide variety of AQG tasks we must specify the task we are interested in to perform our experiments. Regarding the output, we are looking for interrogative English sentences. Having defined the output, we can now move to specify the input and the input/output relationship. As an input we consider English text, where the relation can be described by the following situation: the output question is answered by the input text, that is, the questions are about the input information. In this task the AQG systems, given a text T as an input, generate a question which can be used, for instance, to verify the respondent's knowledge about T. We will give some examples in section 4. For the sake of simplicity, in what follows we will refer to this task as text2text. To our knowledge, at the moment the text2text task is the most studied task in the AQG community.

Gold standard evaluation approach: As a generation task, whose possible output space is very large, text2text faces the problem of the absence of a comprehensive gold standard. A gold standard, usually obtained through human annotations, is thought to be the correct set of solutions for a particular task T, and it is used either for training or evaluated models for T's purpose. Given a task, the gold standard evaluation approach is based on the assumption that we can always define a precise and exhaustive (or near exhaustive) set of outputs for the task.

In NLG the gold standard is a set of natural language words, sentences or more generally texts. Given the huge human ability to use many different expressions to reach the same communicative goal, language generation tasks cannot safely rest on the gold standard approach for evaluation purposes.

In NLP the gold standard approach has been used for evaluation through automatic metrics, which compare the automatic generated sentences against the gold standards. In the last few years many studies in NLP have shed light on the correlation between human judgement and metrics such as BLEU (Papineni et al., 2002)[5]. The results, which have shown this correspondence to be weak[6], cast doubt on the feasibility of use of these metrics to evaluate the overall model quality. Indeed, a model may generate very high quality sentences that are different from the ones in the gold standard set used for evaluation — this is especially possible if we train and evaluate a model with different corpus text.

For the text2text task, this problem can be expressed in the following way: Given a text T it is possible to generate many different correct questions about T. More specifically, given a text T with an answer A relative to T, it is possible to generate different correct questions answered by A. In both cases the questions generated, although good, can be very different from the gold standards previously chosen, and so get a low ranking from automatic metrics. For NLG tasks, the idea of some platonic gold standard which, given a task t, can provide an ideal solution for t, is unrealistic.

As we will see soon, the difficulty of having a platonic gold standard is not just a problem for automatic evaluation metrics but it is also reflected in the difficulty of reaching a high human evaluation agreement.

Annotation guidelines: The absence of a comprehensive gold standard evaluation approach is reflected in the difficulty of formulating precise annotation guidelines. These are either defined in order to annotate corpus data or to give instructions for human evaluation tasks.

The annotation guidelines accompany a particular annotation scheme[7] , and should strictly define the features that the humans have to annotate, as well as how they should be annotated. The goal is that of reducing the annotators' subjective interpretation so as to have a consistent corpus. Indeed, human annotations are prone to subjective bias, so the guidelines have the aim of reducing this subjectivity to make the annotation process more reliable. Although we do not have a standard way to create annotation guidelines, some common ground rules have been defined. Usually a sound annotation guideline introduces at least some criteria which define the annotation task, alongside a description and some examples whose aim is to help the annotators to understand the criteria (see for example, Palmer and Xue (2010) and Pustejovsky and Stubbs (2013)). In a language generation task, where the possible space of outputs cannot be strictly defined, as it is subject to individual human differences, the guideline definition be-

[5]For an exhaustive list of automatic metrics used in NLP we refer to (Gatt and Krahmer, 2017) page 126.

[6]For an in depth discussion about this point we refer again to (Gatt and Krahmer, 2017), especially the section 7.4.1 and the reference there presented.

[7]An annotation scheme determines the conceptual content of the annotation task, for example by identifying the set of legitimate alternatives the annotator can choose from.

comes particularly difficult and there is a high risk of ending up with a set of poorly defined criteria. Far from there being a lack of precision in writing the guideline, here the problem seems to be intrinsic to the nature of human language. A fascinating analogy that aims to explain this point, presented by Geoffrey Sampson in the online page https://www.grsampson.net/RSue.html, is the following:

> Suppose we wanted to be able to say how large particular clouds are – what volume of space they occupy. Clouds are fuzzy things, so one problem would be what we mean by the volume of a cloud – what exactly should we count as its edge? But even if we adopted some precise definition of cloud boundaries, so that it became meaningful to say that this cloud is exactly N cubic yards in size, not N + 1 or N - 1, it might still be beyond mankind's abilities actually to measure clouds so exactly.

As a result of the fuzziness of human language, the criteria defined in an annotation guideline can be quite vague and be left to annotators' interpretation, which can result in a low IAA.

Kappa coefficients for IAA: The IAA measures the agreement between annotators. Thanks to this measurement we can estimate the annotation reliability. Generally, it is believed by the NLP community that the annotations are reliable if annotators agree, to a certain extent, on the category assigned. The kind of extent is determined by the method chosen to measure the agreement. Usually the IAA is performed by Kappa coefficients K[8].

The common theme to a variety of formulations is that K corrects annotators' agreement by the expected chance agreement:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times the annotators agree, whereas $P(E)$ is the proportion of times the annotators would be expected to agree by chance.

Kappa coefficients are used with Kappa scales of interpretation. For example, Krippendorff (1980) considers good any data annotation with agreement in the interval [0.8, 1], tentative any data annotation where agreement is in the interval [0.67, 0.8] and to discard any data annotation where agreement is below 0.67 [9]. Given the arbitrary nature of these choices of number, we can find different scales for K interpretation (see for example Landis and Koch (1977) , Green (1997) and for a general discussion Artstein and Poesio (2008)).

Since Carletta (1996), the K coefficients and its scales interpretation have been introduced in the area of NLP. Carletta took these metrics from content analysis in order to "compare results in a standard way across different coding schemes and experiments and to evaluate current developments" specifying that "whether we have reached (or will be able to reach) reasonable level of agreements in our work as a field remains to be seen".

3 The case of Automatic Question Generation

Problem statement: Human language generation tasks, given the richness of human languages, can usually not be evaluated with respect to some platonic gold standard. Given a sentence, different persons can have conflicting opinions about the sentence's ability to fulfill the task for which it was generated. More generally, given a sentence S, different persons can have conflicting opinions about the fluency, the grammaticality, the naturalness, the elegance etc. of S. These discrepancies are reflected, firstly, in the difficulty of generating strong annotation guidelines, and secondly, in the difficulty of reaching

[8] Here we are following the notation used in Carletta (1996). It is worth notice that the K formulation presented catch the most used agreement metrics from the NLP community, such as the π coefficient of Scott (1955) its Fleiss' generalization (Fleiss, 1971), and the k coefficient of Cohen (1960). The α coefficient of Krippendorff (1980) is expressed in a similar way but in term of disagreement. This difference is not relevant to the aim of this paper, so in what follows when we use the term K coefficient we are referring at some instance of the above mentioned metrics. We refer to (Artstein and Poesio, 2008) for an excellent survey of the several agreement coefficients present in the literature.

[9] Arstein and Poesio (2008) note that: *the description of the 0.67 boundary in Krippendorff (1980) was actually "highly tentative and cautious," and in later work Krippendorff clearly consider 0.8 the absolute minimum value of α to accept for any serious purpose: "Even a cutoff point of α = .800 ... is a pretty low standard"* (page 576).

a satisfactory IAA in the evaluation phase — where what is a satisfactory IAA is defined by some K interpretation scales. Given the nature of the tasks in play, we believe it is time to rethink the IAA in the NLG community. Far from being a problem, a low IAA score reflects the richness, variability and beauty of human languages.

IAA in text2text task: A careful examination of the text2text literature shows that the human evaluation IAA, where reported, is usually below 0.6[10]. This means, following the Krippendorff Kappa interpretation scale, that the evaluations are to be considered inadequate for a satisfactory analysis.

It is also worth noticing the absence of shared annotation guidelines in the area. Indeed, we can find a large variety of criteria used in the human evaluations alongside more or less detailed evaluation guidelines. Although this absence is a deficiency in the area, which makes it hard to have a uniform way to check the quality across generation systems, we cannot completely ascribe to it the low IAA reached in the evaluation phases.

About this point we can find an analysis by Sampson and Babarcz (2008). As we said before, the authors perform their experiment using the SUSANNE scheme (Sampson, 1995), which is an annotation guideline of 449 pages. Of the three conclusions they reach in the study, the most interesting to the aims of this paper is the following: "while it is not easy to define watertight boundaries between the categories of a comprehensive structural annotation scheme, limits on inter-annotator agreement are in practice set more by the difficulty of conforming to a well-defined scheme than by the difficulty of making a scheme well defined".

It is worth noting that Sampson and Babarczy's experiment was done with two experts. Despite the fact that the annotation guidelines were very precise and the annotations were done by experts, divergent rank opinions still emerged from the experiment.

Experiments description: The current paper was motivated by our attempt to define annotation guidelines for the text2text task. The methodology we used was that of refining the criteria chosen through several iterations of discussions and pilot evaluations. During these iterations we noticed that regardless of how many changes we made, there remained a divergence in the judgements that we could not reduce by modifying the guidelines.

We started by choosing the criteria to use. We decided to study the evaluation along two dimensions, a linguistic one, which aimed to evaluate the question quality from a grammatical and idiomatic point of view, and a task-oriented one, which aimed to evaluate how well the questions fulfill the task for which they were generated[11]. In the text2text case the task is that of generating a question in relation to a given paragraph. So in the task-oriented dimension we were interested in the degree to which the question was answered by the input text. Although these two dimensions are tacitly shared between researchers, making them explicit has the merit of outlining a common space to check quality across generation systems. Within this dimension we defined four criteria, which we will present in the next section. We aimed for a sound description of each criterion, accompanied by examples. Once we defined the first version of the annotation guideline we tested it by performing a pilot assessment phase.

We tested our evaluation guidelines taking random input paragraphs and questions from the SQuAD dataset (Rajpurkar et al., 2016). We repeated this process three times. In each iteration we took five paragraphs and for each paragraph six questions. All the questions were human generated. Between these, three were taken from the reference questions of the input paragraph (for one of these we automatically swapped some words in order to change the grammaticality), and three were about the topic of the input paragraph, although generated with a different (but with similar topic) paragraph as an input. We asked two volunteers to join us in the process of annotating the questions. Both volunteers were

[10]To our knowledge the only significant exception is (Godwin and Piwek, 2016), where, as we said before, the authors define an interactive process in which the annotators can discuss their opinions about the criteria used in the evaluation. Such a process presupposes that the annotators talk with each other in order to profitably improve the criteria in play. In this we can see at least two difficulties. On the one hand, this process is time consuming and is not practicable in Crowdsourcing evaluations. On the other hand, although the process allows a small set of people to agree among themselves, we could have different results with different groups of people.

[11]Gatt and Krahmer (2017) note that traditionally, in NLP, the linguistic dimension is defined by the *fluency* or *readability* criteria, whereas the task-oriented one is defined in term of *accuracy*, *adequacy*, *relevance* or *correctness* criteria.

Spanish, who had lived several years in the UK or US, but had not formally studied linguistics. Each pilot was performed with four people: two of the authors of this paper and the two volunteers. At the end of each iteration, with the pilot example support, the authors of the present paper discussed the adequacy of the criteria. Based on that discussion we decided how to improve description of and examples for the criteria.

After several iterations we concluded with a final iteration. This time seven annotators, including the two of the authors of this paper, engaged in the annotation task. Between the seven annotators three were native speakers of English. The other five were proficient in English. Three of the annotators had a background in linguistics. Once again we used the input paragraphs from the SQuAD dataset. This time we took two paragraphs and five question for each paragraph. One of the questions was automatically generated by the question generation algorithm freely available at the page https://kjmazidi.pythonanywhere.com/. Three were human generated questions from the SQuAD dataset: two were taken from the reference questions of the input paragraph (for one of these we automatically swapped some words in order to change the grammaticality), and one was about the topic of the input paragraph, although generated with a different input paragraph. The last question was generated by the first author of this paper, who did not personally take part in the evaluation. At the end of the evaluation, the annotators discussed their responses for each criterion and for each question, one by one. During this discussion, we identified a series of differences. As we will see in the next sections, they were based on annotators' subjective taste and experiences. Again, for each criterion, we got a IAA lower than 0.67, with the lowest score being 0.11.

4 A new taxonomy of divergences

We classify the main sources of disagreement we found through our experiments in five categories: i) *Style and taste*; ii) *background knowledge*; iii) *personal assumptions*; iv) *use of common sense inferences*; v) *attention to detail*.

We are going to present some examples for each category in order to explain them. As we said before, we conducted the study along two dimensions, a linguistic one and a task-oriented one. The annotation criteria attempted to characterise these two dimensions. Following the process we described in the previous section, we attempted to improve the criteria by clarifying their descriptions, adding clarifying examples and splitting them into sub-criteria.

In the last version of our evaluation guidelines, we had the following four criteria: i) Pertinence, ii) Grammaticality, iii) Comprehensibility, iv) Fluency. The Pertinence criterion is for the task-oriented dimension. We used this name to underline the need for the question to be directly and strictly related to the paragraph. The linguistic dimension was split into three sub-dimensions or criteria: Grammaticality, Comprehensibility and Fluency. The grammaticality is intended to check possible grammatical errors in the questions. Comprehensibility aims to better frame the grammaticality judgement. Indeed, it can happen that a question, although ungrammatical, is perfectly understandable[12]. That is, it can be possible to work out what the question is asking, even if it is ungrammatical. Finally, the fluency criterion is intended to judge whether the question is idiomatic/natural. The Pertinence criterion is ranked on a scale from 0 to 3, whereas the other criteria use a binary scale. We decided to evaluate the linguistic dimension without providing the paragraph while the task dimension after reading the paragraph. Indeed, we believe that the grammaticality, comprehensibility and idiomaticity of a question can be judged independently of the paragraph from which it was generated. So each annotation was carried out in two stages. First, the annotator was presented with the question in isolation, and asked to provide a judgement on the questions of grammaticality, comprehensibility and fluency. Next, the annotator was presented with both the question and the input text, and asked to provide a judgement on the pertinence.

In our experiments, of the 100 questions we analysed, we found a total of 55 divergences in the linguistic dimensions and a total of 78 divergences in the task-oriented dimension. We present some examples of these divergences in order to explain how the above categories were delineated.

[12]We discuss some examples of this in the next section.

Style and taste: We found some divergences were related to the annotator's taste and their writing style. This kind of divergence emerged in the question on idiomaticity judgements. For example, we notice that American and British English differences played against question idiomaticity. Given the question:

Jean De Rely's illustrated French-language scriptures were first published in what city?

we found divergent judgements because the question sounded awkward to some British annotators, who preferred the use of "which" rather than "what". Similarly, we had divergent judgements with the question:

The adaptive immune system must distinguish between what types of molecules?

where one of the British annotators not only preferred the use of "which" instead of "what", but also marked the question as not idiomatically natural because he preferred the question written in the reverse order: Which types of molecules must the adaptive immune system distinguish between? In a case like this, we can see how the personal writing style influences the evaluation. This eventuality is a result of the fact that, in a generation task, we can use very different sentences, in this case questions, in order to reach the same communicative goal.

Background knowledge: Another issue we found concerned the amount of background knowledge that is needed in order to understand a question. Where annotators did not share relevant background knowledge, they often gave different judgements for a question's comprehensibility and pertinence. For example the question:

How many Time incarnations can a Lord have?

even if not grammatical (the original question is: How many incarnations can a Time Lord have?), was recognised as comprehensible by the annotators who were aware of the 'Doctor Who' television programme, whereas it was marked as not comprehensible by the ones who did not know the show. In the same way, given the paragraph:

Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which recognize components that are conserved among broad groups of microorganisms...

the question:

What part of the innate immune system identifies microbes and triggers immune response?

raised divergent opinions for two reasons: on the one hand, it was necessary to know that microorganisms are the same as microbes. On the other, it was necessary to know that the pattern recognition receptors are part of the innate immune system. The lack of this knowledge from some annotators, resulted in a divergence in ranking the pertinence criterion. In these examples, we can see how the annotators' knowledge determines the evaluation. Indeed, personal knowledge and experiences are reflected in the way we generate and understand sentences. So, divergences in knowledge can be reflected in evaluation divergences.

Personal assumptions: Other divergences arose from different annotator assumptions. Also in these cases the divergences emerged predominantly in the question's comprehensibility and pertinence judgements. For example the question:

Which team finished the regular season?

was considered not comprehensible by the annotators who assumed that all teams would finish the season. These annotators considered the question meaningless. In contrast, other annotators considered a scenario in which some teams could not finish the season. In this case the question was marked as comprehensible. A similar problem was found with the question:

What was the win/loss ratio in 2015 for the Carolina Panthers during their regular season?

Given the paragraph:

The Panthers finished the regular season with a 15-1 record, and quarterback Cam Newton was named the NFL Most Valuable Player (MVP)...

one annotator noticed that the question can be marked pertinent under the assumption that the Carolina Panthers in the question and the Panthers in the paragraph are referring to the same team. He did not make this assumption. Other annotators did make this assumption and ranked the question as pertinent. In a similar way to background knowledge and experiences, personal assumptions are reflected in the way we generate and understand sentences. Again, divergences in assumptions can be reflected in divergences in evaluation.

Use of common sense inferences: This kind of divergence emerged predominantly in the question pertinence judgements. For example, given the paragraph:

The availability of the Bible in vernacular languages was important to the spread of the Protestant movement and development of the Reformed church in France. The country had a long history of struggles with the papacy by the time the Protestant Reformation finally arrived. Around 1294, a French version of the Scriptures was prepared by the Roman Catholic priest, Guyard de Moulin. A two-volume illustrated folio paraphrase version based on his manuscript, by Jean de Rély, was printed in Paris in 1487.

the question:

Jean De Rely's illustrated French-language scriptures were first published in what city?

was marked as not pertinent by one annotator who noted that the paragraph provides information about the place where Jean De Rely's scriptures were printed and not where they were published. To consider this question as unambiguously answered by the paragraph it is necessary to assume that the place where the Jean De Rely's illustrated scriptures were printed is the same where they were published. Although this inference goes much further than the information presented in the text, other annotators made it and marked the question as pertinent. Likewise, the question:

When did the first French language bible appear?

was marked as pertinent by one annotator, who inferred the answer 1294 from the paragraph. Other annotators considered the question as not pertinent because the answer cannot be correctly inferred from the paragraph. As a matter of fact, in the text there is no mention of the fact that the Guyard de Moulin version of the Scriptures was the first French language bible to appear and at the same time it is not possible to rigorously reach this conclusion by the information provided by the text. Here we faced another problem: people reason in different ways. Obviously this divergence also emerges in the way people generate, understand, and in this case evaluate, sentences.

Attention to detail: We noticed that in some cases, annotators overlooked some question details. Also in these cases, the divergences emerged predominantly in the question pertinence judgements. For example, given the paragraph:

The most impressive examples of rococo architecture are Czapski Palace (1712-1721), Palace of the Four Winds (1730s) and Visitationist Church (façade 1728-1761).

the question:

What type of architecture is the Palace of Four Windows an impressive example of?

was ranked as pertinent by some annotators who did not notice that the question uses the word "Windows" instead of "Winds", but it was ranked as not pertinent by the annotators who did notice this detail. Another example of this kind of rank divergences is the following. Given the paragraph:

Victorian farms produce nearly 90% of Australian pears and third of apples. It is also a leader in stone fruit production. The main vegetable crops include asparagus, broccoli, carrots, potatoes and tomatoes. Last year, 121,200 tonnes of pears and 270,000 tonnes of tomatoes were produced.

the question:

How many tonnes of tomatoes does Victoria produce?

was ranked as pertinent by some annotators who did not notice that the paragraph was speaking about "last year" production, whereas the question is more general (maybe it is asking for an annual average, which is not mentioned in the paragraph). Other annotators did notice this detail and ranked the question as not pertinent. This problem is linked to the fact the people can generate and interpret sentences to different levels of generality and detail. This is reflected in the way people understand sentences, in this case the questions, in the context from which they are generated.

We conclude this section by observing that together with the five sources of disagreement we have just presented, we found other sources of discrepancy between annotators — which we suppose are present in each annotation task — related to distractions, misunderstanding the guidelines, or forgetting how to apply the guidelines.

IAA, a new research direction: Many studies have shown IAA appearing to depend on several factors, for example: data typology, data ambiguity, the number of categories used in the annotation, the number of annotators, the use of expert or non-expert annotators, and annotators' attention, skills, memory and training. In the present study we found that, in a language generation task such as text2text, we also need to add the following factors: the annotators' background knowledge, their personal taste and personal assumptions, their attention to detail and their inferential skills.

The obvious conclusion we can draw is that human annotators' disagreements are part of the nature of language. Nevertheless, much effort has been spent on trying to understand the source of this disagreement in order that the disagreement can be reduced, for example by modifying the annotation scheme. According to our experiments, we believe that in NLG tasks, attempts to create annotation guidelines that strongly reduce disagreement among annotators, are in danger of missing the goal of producing human language variability. Indeed, such methods can artificially restrict the annotators' language interpretation and their opinion about sentences' goodness. Consequently, the methods do not catch the model's ability to generate language variability.

Rather than regarding the evaluation disagreement as a problem to be fixed, we suggest that it should be thought of as an ineliminable feature of generation tasks, which reflects the variety of human languages and its users. To this end we propose that the scales of K interpretation should be reconsidered,

for example by introducing confidence intervals. These would delineate boundaries within which we expect to find the IAA. An evaluation agreement that falls below the low boundary can suggest that the evaluation is too unreliable to be considered further. At the opposite end, an evaluation agreement that exceeds the upper boundary may suggest that the task was overconstrained and therefore possibly of limited interest. Such an interval, recognising that a greater level of disagreement between annotators—so that the level of disagreement falls inside the boundaries — is not problematic, would take into account the phenomena we have introduced in the previous section. Given a corpus to evaluate, different groups of annotators can end up with different levels of disagreement regarding the quality of that corpus. These differences, mainly due to how language is produced and understood, as long as they lie inside the interval boundaries, can be considered as inherent to the task. Although there is not a general way to determine these intervals — because of the range of tasks in NLG, such confidence intervals should be defined on a task by task basis — we believe a first step could be to analyse the correlation between differences in IAA and growing linguistic difficulty. Indeed, consider a situation where a generative system is evaluated on its capability to produce human-like language (e.g. as in a Turing test). In such a scenario, it is likely to get the best agreement in those cases were the generated language is somewhat simplistic (for example very shallow sentences or sentences clearly ungrammatical or completely out of context). Conversely, in the cases where the generated language is more complex from a semantic point of view, as in our experiments, it is likely achieve lower agreement. If, using high levels of linguistic sophistication, a similar level of agreement can be reached, then it is possible to consider that level as a bottom threshold. Similarly, it can be done with low levels of linguistic sophistication in order to choose a top threshold. We are undertaking some experiments in this direction for text2text tasks.

In the same vein, following the Shared Task and Evaluation Campaigns (STECs) proposal (see for example (Gatt and Belz, 2010)), we believe that for each NLG task, a common and shared evaluation guideline should be defined, maybe together with a corpus data used only for evaluation purposes. Furthermore, inside such a framework, researchers could also share, alongside their results, their methodologies and ideas to attempt to understand and classify any divergences between annotators. A better understanding of these divergences can indeed help us to understand and improve the tasks at hand. For each NLG task, a new IAA interpretation, alongside a shared evaluation guidelines, could bring in each NLG community a uniform way to check the quality across generation systems.

In order to better interpret IAA, instead of focusing exclusively on interannotor agreement, more attention may need to be paid to the internal consistency of the responses of each annotator. We have argued that while there are irreducible differences between annotators' judgements, the judgements of an individual annotator should be consistent, if they are to represent that annotator's particular linguistic abilities and preferences.

Conclusion: In this paper we investigated the IAA problem for a text2text task. We performed some experiments that show how a low IAA is inherent to the task of language annotation. Our experiments have outlined that factors such as: annotators' background knowledge, personal taste, personal assumptions, attention to detail and inferential skills are presents in evaluation judgments. We believe that these kind of factors are intimately connected to the people experiences and their language understanding and use, and cannot be effectively removed by the choice of annotation guidelines. We believe that, in generation tasks, the efforts of reaching a high IAA run the danger of leading research and system building away from the goal of generating text that is natural and human-like (and possibly towards artificial non-natural language). Indeed, it could not take into account the differences that make human language so broad in its meaning and use. For this reason, we suggest new research directions, for example the introduction of confidence intervals for the IAA interpretation and attention on internal annotator consistency. In our opinion such a shift in interpretation would take better account of the richness and variability in human language.

Acknowledgements

We warmly thank Erika Renedo Illarregi, Luisa Ruge, German Ruiz Marcos, Suraj Pandey, Simon Cutajar, Neil Smith and Robin Laney for taking part in the experiments and sharing with us opinions and

feedback. We would also thanks Karen Mazidi to give us the login access to her online Question Generator. We finally thanks the anonymous reviewers for their helpful suggestions.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4): 555-596.

Petra S. Bayerl and Karsten I. Paul. 2011. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37(4): 699-725.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2): 249-254.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37-46.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378-382.

Albert Gatt and Anja Belz. 2010. Introducing Shared Tasks to NLG: The TUNA Shared Task Evaluation: In Emiel Krahmer and Mariët Theune. (eds.) *Empirical Methods in Natural Language Generation*, pages 264-293.

Albert Gatt and Emiel Krahmer. 2017. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61: 65-170.

Annette M. Green. 1997. Kappa statistics for multiple raters using categorical classifications. *In Proceedings of the Twenty-Second Annual Conference of SAS Users Group*, San Diego, USA.

Vojtěch Kovář 2016. Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications. *In: Recent Advances in Slavonic Natural Language Processing, RASLAN,* pages 127-134.

Vojtěch Kovář, Miloš Jakubíček and Aleš Horák 2016. On Evaluation of Natural Language Processing Tasks Is Gold Standard Evaluation Methodology a Good Solution? *In: Proceedings of the 8th International Conference on Agents and Artificial Intelligence,* Rome, pages 540-545.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology,* Sage Publications, Beverly Hills, CA.

Keith Godwin and Paul Piwek. 2016. Collecting Reliable Human Judgements on Machine-Generated Language: The Case of the QG-STEC Data. *In Proc. INLG16,* pages 212-216.

Art Graesser, Vasile Rus and Zhiqiang Cai. 2008. Question classification schemes. *Proceedings of the 1st Workshop on Question Generation.*

J. Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics:* 33(1): 159-174.

Hector Martinez Alonso, Anders Johannsen and Barbara Plank. 2016. Supersense tagging with inter-annotator disagreement. *Linguistic Annotation Workshop,* Aug 2016, Berlin, Germany, pages 43-48.

Hwee Tou Ng, Chung Yong Lim and Shou King Foo. 1999. A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. *Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources,* Aug 2016, College Park, MD, USA, pages 9-13.

Jekaterina Novikova, Ondřej Dušek and Verena Rieser. 2018. RankMe: Reliable Human Ratings for Natural Language Generation. *arXiv:1803.05928*.

Martha Palmer and Nianwen Xue. 2010. Linguistic Annotation. In: A. Clark, C. Fox, and S. Lappin. (eds.) *The Handbook of Computational Linguistics and Natural Language Processing,* Wiley-Blackwell, Chichester, pages 238-270.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* Philadelphia, Pennsylvania – July 07 - 12, 2002, pages 311-318.

Paul Piwek and Kristy Elizabeth Boyer. 2012. Varieties of Question Generation: Introduction to this Special Issue. *Dialogue and Discourse,* 3(2): 1-9.

James Pustejovsky and Amber Stubbs. 2013. Natural Language Annotation for Machine Learning. *OReilly Media, Inc.*

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Empirical Methods in Natural Language Processing (EMNLP).*

Ehud Reiter and Somayajulu Sripada. 2002. Should Corpora Texts Be Gold Standards fo NLG? *in: Proceedings of the Second International Conference on Natural Language Generation,* pages 97-104.

Vasile Rus, Zhiqiang Cai and Art Graesser. 2008. Question Generation: Example of A Multi-year Evaluation Campaign. In: V. Rus, and A. Graesser. (eds.) *Online Proceedings of 1st Question Generation Workshop, NSF, Arlington, VA.*

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev and Cristian Moldovan. 2012. A detailed account of the First Question Generation Shared Task Evaluation challenge *Dialogue & Discourse,* 3(2):177-204.

Geoffrey R. Sampson. 1995. English for the Computer: The SUSANNE Corpus and Annotation Scheme. *Oxford: Clarendon Press (Oxford University Press).*

Geoffrey Sampson and Anna Babarczy. 2008. Definitional and human constraints on structural annotation of English. *Natural Language Engineering.* 14(4): 471-494.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly,* 19(3): 321-325.