Is NLP Ready for Standardization?

Anonymous EMNLP submission

Abstract

While standardization is a well-established activity in other scientific fields such as telecommunications, networks or multimedia, in the field of AI and more specifically NLP it is still at its dawn. In this paper, we explore how various aspects of NLP (evaluation, data, tasks...) lack standards and how that can impact science, but also the society, the industry, and regulations. We argue that the numerous initiatives to rationalize the field and establish good practices are only the first step, and developing formal standards remains needed to bring further clarity to NLP research and industry, at a time where this community faces various crises regarding ethics or reproducibility. We thus encourage NLP researchers to contribute to existing and upcoming standardization projects, so that they can express their needs and concerns, while sharing their expertise.

1 Introduction

Most of the Natural Language Processing community remains estranged from standardization. As this is already common practice in many computer science fields, including telecommunications, networks and multimedia, what is making NLP so special in that regard? Zielke (2020) has already asked the question of the potential barriers and benefits of standardization work in the broader field of Artificial Intelligence, which is now becoming a reality. Here we propose to deepen that discussion by investigating the more specific context of NLP and its standardization needs.

Standards are normative documents produced by Standards Developing Organizations (SDOs) such as ISO. In practice, they can be of various nature and content. Some of them are terminological references that establish shared terms and definitions for a technical domain. For instance, the ISO Online Browsing Platform[1] indexes all existing ISO definitions. Other standards rather describe a reference framework, which can prove useful for bootstrapping new activities or rationalizing existing ones. Standards can also provide technical specifications for data, systems or procedures. This notably includes quality specifications, as well as interoperability ones (APIs, protocols, etc.). For instance, the C language (ISO/IEC 9899), the MP3 coding (ISO/IEC 11172-3), the Latin-1 charset (ISO/IEC 8859-1) and the 2-letter language codes (ISO 639-1) are all examples of standards.

Standards are written by volunteer experts from various backgrounds (scientific, legal, standardization experts, etc.). In most SDOs, registration is open to anyone willing to contribute, usually through a mirror committee within their national standards organization. Experts collaborate within working groups and decisions are taken by consensus[2] – across countries, but also across backgrounds, across sectors, across technical fields. This approach makes standardization a rather slow process (with up to 3 years to establish some standards), but it also ensures the strength of the agreements.

Standards are especially important for the industry and for regulatory authorities – but they can apply on very technical fields, including scientific topics, and therefore affect the research community as well. This paper investigates how NLP standardization could impact our community, offering both challenges and opportunities.

After a brief review of the current state of NLP standardization initiatives (§2), we explore standardization gaps pertaining to NLP evaluation (§3), data and formats (§4), tasks (§5) and higher-level

[1] https://www.iso.org/obp

[2] Compared to unanimity, where everyone supports the decision, consensus means that noone objects to the decision. This decision process helps to identify middle ground solutions that everyone in an heterogeneous group can find acceptable, whereas an unanimity requirement would bear the risk of freezing projects due to unsolvable cultural differences or diverging interests.

concepts (§6). Having built a broader view of how NLP standardization could benefit research, but also the society, industry and regulations (§7), we then conclude on possible contributions that NLP researchers could add to those efforts (§8).

2 Existing initiatives towards NLP standardization

Within the NLP community, most of the standardized material is actually *de facto* standards: data, tools or methodology that are consensually used throughout the field, even though they don't have any official status and have not necessarily gone through the formalization process that official standards offer. Such *de facto* standards often result from past shared tasks: for instance, the `mteval-v13a.pl` evaluation script from the WMT shared tasks series has been used for years as the reference evaluation script by a large part of machine translation research. Similarly, the CoNLL-X format for dependency treebanks has been widely adopted following the corresponding CoNLL shared task (Buchholz and Marsi, 2006). As for event detection, the definition of the task itself is fully driven by the ACE 2005 campaign (Walker et al., 2006), to the point that it is sometimes referred to as "*ACE event detection*" (Chen et al., 2018). Recent years have seen however the growth of the Universal Dependencies initiative (Nivre et al., 2016), for establishing common guidelines for treebank annotation across languages; considering the breadth of its contributors and its sustained efforts for guidelines formalization, this project has now become very close to standardization work and could be considered as an SDO.

Regarding official standardization initiatives relevant for NLP, the most established one is ISO's Technical Committee 37 (*Language and Terminology*), and more prominently its subcommittees 4 (*Language resource management*, created in 2001) and 5 (*Translation, interpreting and related technology*, created in 2012). With a strong focus on corpora and annotation, these groups have notably released a number of annotation framework standards (e.g. for TIGER-XML or TEI), which are extensively used by the corresponding industry; they also co-organize with ACL the ISA workshop series on Interoperable Semantic Annotation (Bunt, 2021). Yet this focus on data leaves the algorithmic and evaluation parts of NLP largely unaddressed.

ISO-IEC's Joint Technical Committee 1 (*Information Technology*) has created in 2017 its subcommittee 42 on Artificial Intelligence. This one is more concerned with algorithms, development methodology, and system evaluation, but at the higher level of AI in general and not delving into NLP-specific aspects. To date, its NLP-related activity has focused on defining a few major concepts, such as NLG, question answering, or OCR. Concurrently, other global SDOs such as ITU-T have also explored NLP standardization, but mostly in the context of specific use cases (e.g. ITU-T H.862.5 "*Emotion enabled multimodal user interface based on artificial neural networks*") rather than the NLP field in general. Hence their remains a gap in terms of NLP standardization.

The European counterparts of ISO-IEC (CEN-CENELEC) have thus created in 2021 their own Joint Technical Committee 21 on Artificial Intelligence, with a group dedicated to kickstarting activities on speech and NLP (ad-hoc group 4, *AI systems for human language processing*, on track to become persistent in 2022). This is the first page of a new chapter, roadmaps are being written today. Now is thus the right time to question what can, and should, be standardized within our field.

3 Is NLP evaluation ready for standardization?

The reproducibility crisis that has spread through the field in recent years (Belz et al., 2021b; Lucic et al., 2022) has renewed the community's interest for fair and reliable evaluation. This has led in the 2020s to a blooming of workshops and shared tasks dedicated to evaluation means (whether human evaluation or automated metrics), such as Eval4NLP, HumEval, GEM, or *Benchmarking: Past, Present and Future* (Eger et al., 2020; Belz et al., 2021a; Bosselut et al., 2021; Church et al., 2021). But those concerns are not new, as illustrated by the 4REAL workshops organized in the 2010s (Branco et al., 2016). Even the terminology of reproducibility has been the topic of debate and clarification attempts for many years in the machine learning community (Drummond, 2009). The existence of the LREC conference series itself is a token of that interest for standardized evaluation, with its first edition dedicating a whole workshop to the lack of a shared strategy, definition and infrastructure for system evaluation (ELSE, 1998; McTait and Choukri, 2003).

For domains like human evaluation of NLG, the

need for more consistent practices is clear enough, and Howcroft et al. (2020) have already advocated for producing standards on both the methodology and the terminology, based on their review of 20 years of NLG with conflicting evaluation criteria. Yet here we argue that even cases with seemingly straightforward automated evaluation can suffer today from the lack of standards.

3.1 On defining metrics

One of the challenges of standardized evaluation is to ensure that the metrics used are defined in a way that leaves no place to ambiguity, which in practice is rarely the case in the field. For instance, even the well-known F1-score, despite its very formal definition as the harmonic mean of precision $\left(\frac{TP}{TP+FP}\right)$ and recall $\left(\frac{TP}{TP+FN}\right)$, becomes ambiguous when applied to tasks such as Named Entity Recognition.

A first issue resides in the common practice of casting this chunking task into a sequence labelling task, through BIO-style token-level encoding: `B-...` labels denote the first token of an entity, `I-...` labels other tokens within the entity, and `O` labels other tokens. This raises the question of which objects are considered for true/false positives/negatives: those labels, or the chunks. For instance, with `B-PER I-PER O B-LOC O` as a reference sequence, predicting `B-PER O O B-LOC I-LOC` yields a score of 60 (micro-)F1 if evaluated as a sequence labeling task (67 F1 if `O` is not considered a class) by looking at tokens, but 0 F1 if evaluated as a chunking task by looking at the predicted chunks (here with exact match). While most experienced researchers know to prefer the latter (and know where to get that information), the youngest researchers as well as industry practitioners are not necessarily aware of this implicit rule. Such confusion can in turn lead to incorrect comparison of models, or incorrect reporting of product performance.

Taillé et al. (2020) report other underspecified aspects, such as the criteria to accept true positives (with or without typing, with partial or exact match...), the use of micro- or macro-averaging, or the existing practice to ignore some classes (such as Other or MISC). As they highlight, these issues also propagate to evaluation of relation extraction, and just one of those can already lead to overestimating the results by up to +3 F1 on a widely-used dataset.

3.2 On implementing metrics

Another challenge is the underspecification of implementation details for those evaluation metrics, even when the metric itself has a non-ambiguous definition.

For machine translation, Post (2018) investigates the divergence in scores that can result from different implementations of the BLEU metric, based on diverging choices of parameters and preprocessing (e.g. the maximum n-gram length, the number of references, or user-supplied and/or metric-internal tokenization). He reports up to 1.8 BLEU difference when varying only the tokenization used for scoring, which is actually more than the gains measured for BPE (Sennrich et al., 2016), which was a game changer for neural machine translation.

Such variations in implementation can occur even in cases as seemingly simple as using F1-scores for classification: for instance Belz (2021) compare concurrent reproduction studies of the same text classifier, and report score divergences up to 5.2 F1 due to metric reimplementation.

Another source of implementation divergence is the procedure adopted to deal with invalid outputs (ill-formatted, impossible sequences, etc.). In the case of Named Entity Recognition, Lignos and Kamyab (2020) investigate how different strategies to repair invalid BIO sequences within the scorer can impact the measured F1, a condition which according to Palen-Michel et al. (2021) also affects the gold labels in a number of renowned datasets, and leads to differences up to 3.25 F1 in a realistic scenario. For the BIOES encoding scheme alone, Kroutikov (2019) numbers at least 7776 different strategies that could be adopted to repair invalid label pairs.

3.3 Tooling to the rescue?

As a means to circumvent those pitfalls, Lignos and Kamyab (2020) advocates for never reimplementing evaluation metrics and relying instead on third-party reference tools. This is in line with Post (2018)'s strategy to release the SacreBLEU package, with the hope that its configurability, documentation, ease of use and variant reporting will enable standardized evaluation. Can tools alone indeed fill in for standards?

The main issue with that view is that it supposes that tools are correctly used. However, Marie et al. (2021) unveil that the growing number of users of SacreBLEU are in practice often misusing it (not re-

porting the variant used, comparing its scores with other scorers, etc.). Similarly, Palen-Michel et al. (2021) release SeqScore as a possible reference tool for named entity recognition evaluation, but they do so based on the failure of previous *de facto* standard tools. For instance, Akbik et al. (2019) observe that their previous paper (which has now over 1000 citations) had overestimated its results by up to 0.8 F1, because they used the official CoNLL-03 evaluation script (designed for BIO) on a BIOES-encoded dataset. On a side note, it can also happen that the most popular scorer simply contains a bug – how can this be assessed if the tool itself serves as standard?

Another possible approach would be to rely more heavily on Kaggle-style benchmarking platforms that enable fairer comparison than standalone evaluation tools. The issue here is that such practices can arbitrarily foster inadequate evaluation. Bowman and Dahl (2021) now consider NLU evaluation "broken" due to benchmark-driven standardization of practices: a number of those benchmarks are actually rewarding unreliable and biased systems. They leave no place to reflect upon a given system's appropriate evaluation setting, and instead incentivize gaming the numbers. Church and Hestness (2019) review 25 years of evaluation practices and show how the rigour efforts that have led to such benchmarks are now pushing against their initial purpose of bringing more insights to "content-free debates". Overall, leaderboards have drawn a lot of criticism in recent years (Rogers, 2019; Ethayarajh and Jurafsky, 2020; Kiela et al., 2021) and are therefore a poor candidate to address the lack of standards.

Instead of producing and relying on tools, typical standardization work would rather approach the issue by writing comprehensive specifications of the evaluation metrics (detailing their computation, their usage, their meaning), which can in turn apply on tools. This includes providing the means to verify that a given scorer or a given evaluation protocol is compliant with the specification. Hence comparable evaluation can be formally ensured, but not at the cost of insights and appropriateness.

3.4 Does it matter?

So the lack of standards leads to more imprecision in the measures and less rigorous comparisons. Is that really an issue, as long as those numbers are high and continue increasing, whatever the criteria? Haven't experimental sciences handled imprecision for centuries, and accepted that challenge as part of the job?

According to Morey et al. (2017), such imprecision has already endangered scientific progress in whole fields of NLP: in their review of several years of contributions in discourse parsing, they discover that the various conclusions drawn on the benefits of distributed representations are mostly wrong in that field. What was considered a huge improvement, with 24 to 51% relative error reduction depending on the metric, was actually a gain of 11 and 16% for two of the metrics, and a *loss* of 15 and 53% for the other two. Here the culprit was the choice to macro-average over documents in some but not all of the works, following practices existing in different communities.

The lack of standards can thus lead to misinterpreting regress as progress. But it can also affect the wider world outside of research. For instance when a contract is signed, and B2B products are to be developed according to a given performance level specified in the contract, there should be no place to ambiguity. Who should be the judge of whether the contract is fulfilled, if the bar is met by one implementation variant of the metric, but not the other? And what if a regulation contains such performance requirement?

Comparability is also a strong enabler for individual rights as consumers. Potential users should be able to make an informed choice when comparing existing products. Transparency regulations can contribute to that, but that information becomes meaningless if the same number can be interpreted differently depending on implementation details.

3.5 Can standards hinder research?

Scientific concerns regarding NLP evaluation go in fact way beyond the need for fair comparison. A number of automated metrics in wide use today have poor correlation with human judgment, and a lot of research efforts have been devoted to designing more relevant metrics. Notoriously, WMT has been running an annual shared task on machine translation evaluation metrics since 2008 (Bojar et al., 2016; Mathur et al., 2020; Freitag et al., 2021), thereby consolidating the community consensus that the BLEU metric certainly has its utility, but also a number of shortcomings (Reiter, 2018), and it is far from being the best metric in existence. METEOR, chrF, CharacTer, BERTscore

(Banerjee and Lavie, 2005; Popović, 2015; Wang et al., 2016; Zhang et al., 2019) are just a few examples among a broad panel of often more appropriate metrics, even though the single-best "one-size-fits-all" metric has not been found yet.

One possible fear with standardization could then be to prevent researchers from pursuing their quest for the best metric, or simply to prevent them from using in their work another good metric instead of BLEU – leading again to fostering bad evaluation practices and limiting the insights brought to future research. However, standards do not need to be compulsory. It is quite possible to write them in a way that preserves that research freedom, but still brings some order and clarity. BLEU is not the best metric, but BLEU is nevertheless preferable to exotic approaches such as measuring an F1-score at the sentence level (true positive if the sentence is an exact match). Are we confident that all practitioners that may have to evaluate a machine translation system at some point (including e.g. software developers in the industry) are aware of that? Can we at least give formal existence to that tiny piece of knowledge?

It is indeed a fact that in a number of cases in NLP evaluation, it is not necessarily known what is the most appropriate choice among the various existing variants. It can also be use case-dependent. And standards in such context are not meant to arbitrarily foster one option among the others. Their role here would rather be to formally reference and specify the existing relevant options (pushing away the ones that are already known to be inappropriate), and offer practical ways to declare, identify or verify which one of those options has indeed been used in a given paper or product.

4 Are NLP data and formats ready for standardization?

Yes they are, and they have been as early as 1993, when EAGLES (Expert Advisory Group for Language Engineering Standards) was established to develop such standards. Ide et al. (2017) review 30 years of community progress from confusion to *de facto* standards to standards. However, despite marked efforts from ISO's Technical Committee 37, this paradigm has only been adopted so far in some parts of the field, and much progress remains to achieve for fully standardizing NLP annotations.

In particular, Ide et al. (2017) underline the need to better standardize the *content* of annotations. While many (although not all) corpus authors have gone through the formalization process of writing annotation guidelines, this has mostly led to a profusion of co-existing guidelines for the same task. The case of dependency parsing is interesting in that regard, as the Universal Dependencies project managed to unify most of the pre-existing annotation schemes, while preserving their idiosyncrasies (Nivre et al., 2016). Yet this is a success story that most parts of the field have not had so far.

In addition, annotation processes should include some quality control mechanisms, such as measuring inter-annotator agreement (Hovy and Lavid, 2010). However, there is poor consensus on what would be a "good" agreement value for a given task, depending on its complexity and subjectivity (Artstein and Poesio, 2008; Mathet et al., 2012). Are we even sure that inter-annotator agreement is an appropriate quality control (Wong and Lee, 2013; Passonneau and Carpenter, 2014; Plank et al., 2014; Boguslav and Cohen, 2017; Basile et al., 2021)? In recent years, the growing reliance on crowdsourcing has only strengthened the challenges, hence the pressing need for standardizing practices (Sabou et al., 2014).

Standardization gaps do not concern solely the semantics of the annotation, but also their format. Looking at machine translation, parallel corpus formats include SGML (for which WMT maintains a `wrap-xml.perl` script to preserve compatibility with scoring scripts), XML (with XCES for sentence alignment), TMX, bitext (two files with corresponding line numbers), but also tabular formats with per-language columns separated by either tabs or other separators. The OpusTools converters (Aulamo et al., 2020) support only part of that spectrum. As for named entity recognition, coexisting encoding schemes include IO, IOB (aka IOB1), BIO (aka IOB2), BIOES (aka IOBES), BILOU (aka BILUO) and BMEOW (Palen-Michel et al., 2021). And there are others (Malik and Sarwar, 2016). One can always write converters, but this is tedious work, and prone to introducing discrepancies in case of invalid sequences (see §3.2). Third-party open source converters can help (Lester, 2020), yet they usually support only some of the encoding schemes. Formats can further differ when considering the file format: whereas CoNLL-2003 was distributed as tabular IOB (Tjong Kim Sang and De Meulder, 2003), spaCy relies on JSON BILUO. And this is only for

sequence tag schemes, while MUC-6 uses SGML (Grishman and Sundheim, 1996) and WiNER-fr prefers an offset-based scheme to directly encode the spans (Dupont, 2019).

In terms of input and output formats, NLP tools can already rely on a number of extensible pipelines such as Stanford CoreNLP or spaCy (Manning et al., 2014; Honnibal and Montani, 2017), as well as abstraction frameworks such as AllenNLP or PyText (Gardner et al., 2018; Aly et al., 2018) – but this differs from actual APIs designed for interoperability *among products*. Today such interoperability is mostly fostered by infrastructure-based initiatives such as the Language Application Grid (Ide et al., 2016, 2015). The European Language Grid project (Rehm et al., 2020a, 2021) now proposes to build an umbrella platform that hosts resources but also unifies NLP APIs through its "functional services" infrastructure. In addition, Kim et al. (2020) propose to standardize a web protocol for NLPaaS, while Rehm et al. (2020b) set a roadmap of interoperability levels to enable cross-platform workflows. Instead of duplicating those projects, the role of SDOs here would rather be to build upon those APIs, by escalating them into official standards with formal specifications.

Finally, data warrants data documentation. This is another area where individual initiatives have produced valuable guidelines on necessary metadata (Bender and Friedman, 2018). But work still remains to give that material more formalism and ensure consensus across communities.

5 Are NLP tasks ready for standardization?

Getting to the core of NLP, even the tasks themselves warrant further consideration for standardization. Indeed, NLP research has recently gained awareness that making further progress on NLU tasks now meant taking some detours to better define terms like "meaning" (Bender and Koller, 2020), "comprehension" (Dunietz et al., 2020) and the associated tasks. Yet even the basic expectations on inputs/outputs can be underspecified for some tasks. For instance, question answering can refer to various concrete tasks, such as multiple-choice answer selection (Aydin et al., 2014), span extraction (with or without paragraph retrieval) in the SQuAD style (Rajpurkar et al., 2016), free-form answering that can include multi-hop questions (Chen et al., 2019), or answering questions over knowledge bases (Fu et al., 2020), which don't warrant the same algorithmic approaches. Gardner et al. (2019) propose to solve the conundrum by considering question answering as a format and splitting it from the definition of the task; yet even then the taxonomy remains dense (Rogers et al., 2021).

Information extraction is another field where tasks and their terminology are largely ill-defined. Even its primary task, named entity recognition, has been subject to a number of conceptual debates (Marrero et al., 2013). Entity linking is better delineated, but has been associated with a number of different names: entity linking, named entity linking, named entity disambiguation, named entity normalization... Are all of those terms synonymous, or do they slightly differ in scope? The literature has already proposed many definitions for entity linking, often inconsistently: for instance Shen et al. (2014) write both "*Entity linking is the task to link entity mentions in text with their corresponding entities in a knowledge base*" and "*to link named entity mentions appearing in web text with their corresponding entities in a knowledge base, which is called entity linking*". This notably raises doubts as to whether entity linking applies only to named entities, or also to non-named entities (Paris and Suchanek, 2021). Or to non-named mentions of entities that have names? In the lack of terminological standards, presumably the best definition of the task is to look at how the corpus at hand has been annotated; but then the task definition can vary a lot from one dataset to another, so that evaluating an entity linking approach on multiple datasets may not make actual sense. Many other discrepancies could be listed here (e.g. relation extraction referring to either relation clustering or open information extraction), but in the end, the name "information extraction" itself is an ill-defined term, with a functional scope that varies a lot depending on individuals. So if a system is branded as an information extraction system, what are its functionalities supposed to be?

Time is not innocent in those terminological conflicts. Language modelling is one striking example of terminological drift. Historically, language models meant "*a probability distribution over all possible word strings in a language*" (Arisoy et al., 2012) – or even a next-word predictor, as in the n-gram paradigm: "*language modeling, the problem*

of predicting the next word based on words already seen before" (Xu and Jelinek, 2004). But since 2018 and the advent of *masked* language models, the term "language model" has now shifted to refer to Transformer-based contextualized embeddings, regardless of any probability distribution, and not necessarily autoregressive (as in Ettinger, 2020).

Are these discrepancies an issue? Semantic drift is a natural phenomenon in any language, and a profusion of definitions also means a profusion of problems addressed by the community as a whole. However, trouble arises when using those task definitions to catalog or to assess existing systems: how to decide whether a given system meet one's expectations, if it is branded with ambiguous functionalities? Achieving clarity on product capabilities is a matter of commercial interest for companies, and of consumer rights for individuals. But it can also affect scientific processes, as exemplified by the Great Misalignment Problem (Hämäläinen and Alnajjar, 2021) between blurry objectives, the actual task fulfilled by the system, and the task against which human evaluation is performed.

6 Are NLP concepts ready for standardization?

At a higher level, a number of concepts would also benefit from formal standards. This notably concerns the term "multilingual", which has been used to describe very different properties, such as: a system that juxtaposes models for multiple languages (Otero and González, 2012), with or without internal language identification; an algorithm that does not rely on language-specific features or knowledge, and can therefore be trained on a dataset from any language (Johansson and Nugues, 2006; Szarvas et al., 2006), even though this does not guarantee actual language independence (Bender, 2011); or a single model that can indiscriminately process contents from many languages (Pires et al., 2019). Focusing only on the latter definition, how many is many? And how diverse? Can an Indo-European-only system be considered multilingual? In light of rising initiatives for fostering more language diversity in NLP research (Bender, 2019; Joshi et al., 2020), including a dedicated theme track at ACL 2022, it now appears pressing to establish consensual criteria on what renders a given system multilingual. Otherwise, how can progress in that matter be quantified?

Trustworthiness is another relevant concept for NLP systems, especially from the viewpoint of policy makers. The High-Level Expert Group on AI (2019) has notoriously established a list of AI trustworthiness characteristics, but they still lack shared actionable definitions. The concept of bias for instance, while subject to a growing interest in NLP research, is rarely formally defined in that literature, or with diverging senses (Blodgett et al., 2020), even though that conceptualization should be a prerequisite before defining the corresponding bias measures (Dev et al., 2021). "Robustness" is similarly overloaded, with meanings ranging from maintained performance on out-of-domain data (Bernier-Colborne and Langlais, 2020), on transformed data (Sanchez et al., 2018; Gan and Ng, 2019), or in presence of natural noise (Zhou et al., 2019), to specific defenses against adversarial attacks (Hsieh et al., 2019). A fortiori, there is no formal taxonomy on what kind of noise a "robust" NLP system should minimally handle: typos only, or L2 learners grammar errors, lexical borrowings? Broken encoding? Or others?

Regarding interpretability and explainability, while some use those terms interchangeably, others have drawn firm distinctions: interpretability is "*loosely defined as the science of comprehending what a model did (or might have done)*" (Gilpin et al., 2018), while "*Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand*" (Arrieta et al., 2020), thereby putting the cognitive load of that understanding process more on the model and less on the human. As it seems that "interpretability" has become the preferred term in the NLP community, while other fields rather use "explainability", should that difference of focus be understood as a conceptual divergence of interests (whereby the NLP community would foster more involvement of the human in model understanding than other communities), or only as a terminological discrepancy? Concurrently, "explainability" as expressed by some other audiences (especially non-technical ones) has nothing to do with either of those concepts, and is rather a synonym for "transparency", "testing", or even "reproducibility" (Brennen, 2020), hence a looming crisis if policy makers mean one while practitioners understand the other. As for transparency, while there is consensus on the need for auditability and documentation, the question of what has to be documented and how is still open (Saxon et al., 2021).

7 On the benefits of NLP standardization

Looking at the long-term impact of the Universal Dependencies project hints at how standardizing NLP could more generally benefit the field and its dynamics. The immediate benefit was a more faithful evaluation across languages, that enabled deeper investigation of cross-lingual transfer. But the existence of common guidelines also created a community incentive to produce more data, by providing the guidance and means for easier extension to dozens of zero-resourced languages. The creation of the project itself opened new fora for community-level collaboration and sharing, thereby giving dependency research a new boost. This has been an opportunity to reflect upon research practices, fostering more systematic studies on annotation scheme impact, better highlighting the gaps in linguistic coverage, and uncovering biases in our view of syntax. Overall, standardization contributes to better driving research, both at the individual and institutional levels. A clear taxonomy makes it easier to identify scientific gaps, more compatible resources and tools offer richer experimental means, and shared definitions guarantee that we are all pushing in the same direction.

Standards also support community-wide adoption of good practices. Even if abundantly discussed and well documented as checklists, especially regarding evaluation (Ribeiro et al., 2020; van der Lee et al., 2019; Gehrmann et al., 2022; Marie et al., 2021; Escartín et al., 2021) and documentation (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019; Ligozat and Luccioni, 2021; Wilkinson et al., 2016), consensual good practices guidelines are not necessarily implemented in practice, in research and even more crucially in industry. Escalating them to formal standards makes it easier to enforce them.

Comparability is another clear benefit for NLP researchers, but even more so for users and consumers. Interoperability can facilitate putting a researcher's ideas into users' hands, with easier integration into products. But it is also a matter of survival for SMEs, for packaging and distributing their products in a competitive environment where Big Tech standalone solutions dominate the market and SMEs struggle to propose large-scale alternatives.

Last but not least, standardizing NLP concepts is a necessary step to refine a shared roadmap to address ethical considerations in NLP (Hovy and Spruit, 2016; Leidner and Plachouras, 2017) – but it is also a prerequisite to regulating abuses and enforcing safe use of NLP that preserves individual rights. At a time where the EU is establishing its AI Act, it has started collaborating with CEN-CENELEC to bring clarity on the terminology and processes needed to legally ensure key trust characteristics, such as robustness, transparency or fairness.

8 Last words: are *we* ready?

This review of various aspects of the NLP ecosystem has shown how the lack of standards can cause confusion, inefficiency, and sometimes even render research efforts detrimental to scientific progress, through misinterpretations and fostering of bad practices. Building and formalizing consensus on key practices and concepts would enable instead a more reproducible, more insightful, more industry-ready and more ethical science.

Admittedly, not everything can be readily standardized: sometimes the scientific material to do so does not even exist yet. And sometimes research freedom and creativity need to be preserved by maintaining concurrent options. But these are precisely cases where it is even more important that the standardization ecosystem benefits from scientific expertise, in order to avoid over-standardizing the field, or widening the discrepancies between research and industry practices.

We believe it is a matter of scientific responsibility to offer such guidance to those who are shaping the industrial and legal future of society-wide use of NLP. Contributing to standardization means sharing our expertise and insights, but also our needs and our concerns, both as scientists and as citizens. NLP is ready and in need – now *we* have to get ready.

There are numerous ways to taking part. While community-internal initiatives should be pursued and fostered, we also encourage European researchers to join CEN-CENELEC/JTC 21 for contributing to its budding roadmap, and worldwide researchers to both pursue resource standardization efforts within ISO/TC 37 and help ISO-IEC/JTC 1/SC 42 to deepen debates that have still only scratched the surface of the upcoming work. Organize events, discuss, share, debate, draft, brainstorm, publish. And NLP standards will be within reach.

Limitations

A significant part of this paper has a purely illustrative value, and the provided set of examples does not convey a comprehensive view of the existing standardization issues. Similarly, despite extensive search, we offer no guarantee of exhaustivity in our inventory of NLP standardization groups, in particular for non-cited SDOs (e.g. IEEE).

The review and discussion are also biased towards a number of European concerns and initiatives, which may be either a symptom of its pioneering position on the topic, or merely a lack of depth in our survey of local initiatives in other parts of the world. National-level standardization efforts are not discussed either.

Finally, this work only scratches the surface of discussing the scientific and industrial feasability of standardization for each part of the field, which may significantly vary from one task or concept to another, depending on their maturity and history.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahmed Aly, Kushal Lakhotia, Shicong Zhao, Mrinal Mohit, Barlas Oguz, Abhinav Arora, Sonal Gupta, Christopher Dewan, Stef Nelson-Lindall, and Rushin Shah. 2018. Pytext: A seamless path from nlp research to production. *arXiv preprint arXiv:1812.08729*.

Ebru Arisoy, Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramabhadran. 2012. Deep neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 20–28, Montréal, Canada. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and parallel corpus diagnostics. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3782–3789, Marseille, France. European Language Resources Association.

Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, and Murat Demirbas. 2014. Crowdsourcing for multiple-choice question answering. In *AAAI*, pages 2946–2953. Citeseer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Anya Belz. 2021. Quantifying reproducibility in nlp and ml. *arXiv preprint arXiv:2109.01211*.

Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors. 2021a. *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Association for Computational Linguistics, Online.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021b. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Gabriel Bernier-Colborne and Phillippe Langlais. 2020. HardEval: Focusing on challenging tokens to assess robustness of NER. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1704–1711, Marseille, France. European Language Resources Association.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. In *MEDINFO 2017: Precision Healthcare through Informatics*, pages 298–302. IOS Press.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of wmt evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Antoine Bosselut, Esin Durmus, Varun Prashant Gangal, Sebastian Gehrmann, Yacine Jernite, Laura Perez-Beltrachini, Samira Shaikh, and Wei Xu, editors. 2021. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics, Online.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

António Branco, Nicoletta Calzolari, and Khalid Choukri. 2016. Workshop on research results reproducibility and resources citation in science and technology of language. *European Language Resources Association*.

Andrea Brennen. 2020. What do people really want when they say they want" explainable ai?" we asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Harry Bunt, editor. 2021. *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Association for Computational Linguistics, Groningen, The Netherlands (online).

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1267–1276, Brussels, Belgium. Association for Computational Linguistics.

Kenneth Church, Mark Liberman, and Valia Kordoni, editors. 2021. *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Association for Computational Linguistics, Online.

Kenneth Ward Church and Joel Hestness. 2019. A survey of 25 years of evaluation. *Natural Language Engineering*, 25(6):753–767.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Nanyun Peng, and Kai-Wei Chang. 2021. What do bias measures measure? *arXiv preprint arXiv:2108.03362*.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, volume 1. Citeseer.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Yoann Dupont. 2019. Un corpus libre, évolutif et versionné en entités nommées du français (a free, evolving and versioned french named entity recognition corpus). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 437–446, Toulouse, France. ATALA.

Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors. 2020. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, Online.

ELSE. 1998. Towards a european evaluation infrastructure for nl and speech. *Workshop at LREC*.

Carla Parra Escartín, Teresa Lynn, Joss Moorkens, and Jane Dunne. 2021. Towards transparency in nlp shared tasks. *arXiv preprint arXiv:2105.05020*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, and Jian Sun. 2020. A survey on complex question answering over knowledge base: Recent advances and challenges. *arXiv preprint arXiv:2007.13069*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Ralph Grishman and Beth Sundheim. 1996. Design of the muc-6 evaluation. Technical report, NEW YORK UNIV NY DEPT OF COMPUTER SCIENCE.

Mika Hämäläinen and Khalid Alnajjar. 2021. The great misalignment problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 69–74, Online. Association for Computational Linguistics.

High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy ai.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, Florence, Italy. Association for Computational Linguistics.

Nancy Ide, Nicoletta Calzolari, Judith Eckle-Kohler, Dafydd Gibbon, Sebastian Hellmann, Kiyong Lee, Joakim Nivre, and Laurent Romary. 2017. Community standards for linguistically-annotated resources. In *Handbook of linguistic annotation*, pages 113–165. Springer.

Nancy Ide, Keith Suderman, James Pustejovsky, Marc Verhagen, and Christopher Cieri. 2016. The language application grid and galaxy. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 457–462, Portorož, Slovenia. European Language Resources Association (ELRA).

Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. 2015. The language application grid web service exchange vocabulary. In *International Workshop on Worldwide Language Service Infrastructure*, pages 18–32. Springer.

Richard Johansson and Pierre Nugues. 2006. Investigating multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 206–210, New York City. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Jin-Dong Kim, Nancy Ide, and Keith Suderman. 2020. Towards standardization of web service protocols for NLPaaS. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 59–65, Marseille, France. European Language Resources Association.

Mike Kroutikov. 2019. 7776 ways to compute f1 for an ner task.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Brian Lester. 2020. iobes: Library for span level processing. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 115–119, Online. Association for Computational Linguistics.

Constantine Lignos and Marjan Kamyab. 2020. If you build your own NER scorer, non-replicable results will come. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 94–99, Online. Association for Computational Linguistics.

Anne-Laure Ligozat and Sasha Luccioni. 2021. A practical guide to quantifying carbon emissions for machine learning researchers and practitioners. Technical report, MILA; LISN.

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, and Robert Stojnic. 2022. Towards reproducible machine learning research in natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 7–11, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Kamran Malik and Syed Mansoor Sarwar. 2016. Named entity recognition system for postpositional languages: urdu as a case study. *International Journal of Advanced Computer Science and Applications*, 7(10).

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Yann Mathet, Antoine Widlöcher, Karën Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset, and Pierre Zweigenbaum. 2012. Manual corpus annotation: Giving meaning to the evaluation metrics. In *Proceedings of COLING 2012: Posters*, pages 809–818, Mumbai, India. The COLING 2012 Organizing Committee.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Kevin McTait and Khalid Choukri. 2003. Setting up an evaluation infrastructure for human language technologies in Europe. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, page 7377, Columbus, Ohio. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Pablo Gamallo Otero and Isaac González. 2012. Deppattern: a multilingual dependency parser. In *Proceedings of PROPOR*. Citeseer.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pierre-Henri Paris and Fabian Suchanek. 2021. Non-named entities–the silent majority. In *European Semantic Web Conference*, pages 131–135. Springer.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiljevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. 2020a. European language grid: An overview. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France. European Language Resources Association.

Georg Rehm, Dimitris Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julian Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John Philip McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš. 2020b. Towards an interoperable ecosystem of AI and LT platforms: A roadmap for the implementation of different levels of interoperability. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 96–107, Marseille, France. European Language Resources Association.

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiljevs, Gerhard Backfried, Jose Manuel Gomez-Perez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou,

Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlova, Dusan Varis, Lukas Kacena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Julija Melnika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. 2021. European language grid: A joint platform for the European language technology community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 221–230, Online. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers. 2019. How the transformers broke nlp leaderboards. *Posted on the Hacking Semantics blog: https://hackingsemantics.xyz/2019/leaderboards*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Saxon, Sharon Levy, Xinyi Wang, Alon Albalak, and William Yang Wang. 2021. Modeling disclosive transparency in NLP application descriptions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2037, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

György Szarvas, Richárd Farkas, and András Kocsor. 2006. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In *International Conference on Discovery Science*, pages 267–278. Springer.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Billy T.M. Wong and Sophia Y.M. Lee. 2013. Annotating legitimate disagreement in corpus construction. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 51–57, Nagoya, Japan. Asian Federation of Natural Language Processing.

Peng Xu and Frederick Jelinek. 2004. Random forests in language modelin. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 325–332, Barcelona, Spain. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 565–571, Florence, Italy. Association for Computational Linguistics.

Thomas Zielke. 2020. Is artificial intelligence ready for standardization? In *European Conference on Software Process Improvement*, pages 259–274. Springer.