16. Assessing Natural Language Processing

Yvette Graham, Trinity College Dublin

This chapter details evaluation techniques in Natural Language Processing, a challenging sub-discipline of artificial intelligence (AI). It highlights proven methods to provide both fair and replicable results for evaluation of system performance, as well as methods of longitudinal evaluation and comparison with human performance. It recaps pitfalls to avoid in applying techniques to new areas. In addition to direct measurement and comparison of system and human performance for individual tasks, the chapter reflects on the degree of shared human-machine task, scalability and potential for malicious application. Finally, it discusses the applicability of human intelligence tests to AI systems and summarises considerations for devising a general framework for assessing AI and robotics.

Introduction

Artificial intelligence (AI) and robotics aim to automate tasks that would otherwise require human intelligence and/or physical ability to complete. The Future Skills Expert Meeting aimed to devise and compile skills and tests suitable for assessing AI and robotics. The best place to look for appropriate tests is within each sub-discipline itself. Within each sub-discipline, tests for experimental procedures provide evidence that a proposed method or approach for improving system performance works or is worthwhile.

This chapter looks at how Natural Language Processing, for example, has developed such tests. Analysis of emerging methods shows that valid and reliable measurements have already been developed for some areas. This include methods of comparison with human performance, as well as longitudinal evaluation of systems. It also includes several examples where evaluation has resulted in inaccurate conclusions to demonstrate the challenge of getting it right.

Taking inspiration from successful methods of Natural Language Processing evaluation, the chapter identifies considerations for constructing a general-purpose framework for measuring system performance in AI and robotics. Such a framework would allow measurement of improvements to systems over time and comparison of system and human performance.

Furthermore, it identifies the three most important factors beyond individual system performance for predicting or measuring the impact of a given AI technology on society and the future workforce. First, a framework should quantify the human effort saved by an AI system by automating an individual task. This would consider if the approach were unsupervised or supervised (and any human effort involved in labelling data). Second, it should quantify the human effort saved by a given AI system when applied at scale. This would consider the feasibility of substantially growing the deployment of a given technology. Third, it should quantify the human effort in monitoring and retraining the technology. In addition, it should identify the potential impact of the technology on society: do most people think it is safe and desirable?

In all of the above, the framework should provide the context in which the results of a given test or measurement apply. The degree to which results apply outside the testing context cannot be assumed. AI technologies, when applied to a new domain, will inevitably require varying retraining and produce distinct results.

Finally, the chapter considers the importance of the AI system's ability to operate independent of humans. Rarely will an AI system fully replace human workers. Rather, AI technologies are much more likely to substantially enhance the way in which humans complete tasks. In some cases, they will vastly increase both the efficiency and scale by which task completion will be possible.

To that end, the chapter outlines a potential method of measuring this dimension of AI system performance, as opposed to pitting human against machine in tests. This attempts to measure the amount of human effort saved in a hybrid human-machine setting.

Existing methods

The vast majority of AI research aims to automate the completion of individual tasks and tests. Methods are subsequently designed to evaluate the system with respect to specific individual tasks. In Natural Language Processing, for example, the research area is generally divided into the following sub-areas with individual tasks evaluated with appropriate methodology for that task:[1]

- Dialogue and interactive systems: development of systems capable of engaging with humans through natural language.

- Discourse and pragmatics: the study of language in its context of use, with pragmatics focusing on the effects of context on meaning, and discourse analysis focusing on written and spoken language and social context.
- Natural language generation: automated creation of natural language text or speech from natural language and/or abstract representations.
- Information extraction: automated retrieval of information from text or speech.
- Information retrieval: obtaining information or resources relevant to user information need.
- Text mining: automatic derivation of high-quality information from text or speech data.
- Language grounding to vision and robotics: techniques for linking language to objects, actions, sounds, images and so on.
- Cognitive modelling: computer science that deals with simulating human problem solving and mental processing in a computerised model.
- Psycholinguistics: the study of the mental aspects of language and speech.
- Machine translation: automated translation of text or speech from one natural language to another.
- Phonology: the study of the patterns of sounds in a language and across languages.
- Morphology and word segmentation: the study of words, how words are formed and the relationship of words to other words in a single language.
- Question answering: automatic retrieval or generation of answers to questions posed in natural language text or speech.
- Semantics: Lexical, Sentence level and Textual Inference: study of the meaning of natural language text or speech and what can be logically inferred from it.
- Sentiment analysis, stylistic analysis and argument mining: automatic classification of the sentiment, style or argument encoded in natural language text or speech.
- Speech and multimodality: automatic processing of spoken and multimodal (image, video, etc.) data.
- Summarisation: automatic extraction of information from natural language text or speech rendered as more concise text or speech.
- Syntax: analysis of the underlying grammatical structure of natural language text or speech (tagging, chunking and parsing).

Machine translation leads the way

In Natural Language Processing, machine translation has led the way in terms of rigorous evaluation methodology. Its evaluation methods that aim to provide a realistic reflection of system performance were first established and later adapted to other Natural Language Processing areas.[2] In machine translation, benchmark-shared tasks in which multiple research teams compete with each other are evaluated. A win at this task identifies the team as a world leader in this research area (Barrault et al., 2020[1]).

Teams are provided with a set of test documents, freshly sourced on line and unseen by participants in the task to translate automatically with systems. In blind tests, large numbers of human assessors provide quality judgements of translations produced by each system. Finally, human evaluation quality ratings are combined into a meaningful statistic/overall score for a given system. The best performing system(s) is identified, considering statistical significance when small differences occur between results for systems.[3]

Figure 16.1. A German test sentence translated by machines and humans

		Rating
Source :	*Im Ziel warf er sein Paddel vor Freude weg und reckte beide Arme siegessicher in die Hohe - wohlwissend, " dass es mindestens fur eine Medaille reichen würde.*	
Machine :	At the finish, he threw away his paddle for joy and raised both arms in victory - knowing that it would be enough for at least one medal.	23.4
Human :	He threw his paddle with joy at the finishing line and, confident of victory, threw both arms in the air - safe in the knowledge that his efforts would secure him a medal.	67.5

Source: Graham et al. (2020[2]).

Tests to evaluate machine translation systems are designed to provide a realistic and fair ranking of systems.

First, to help provide realistic results, test data used in translation are freshly sourced and unseen by systems. This is akin to a realistic use case where the input text to be translated will certainly neither be known nor predictable to systems. Figure 16.1 provides example test data freshly sourced from an online news article and used in a past machine translation competition.

Second, human assessment is employed as opposed to an automatic metric of some description. Automatic metrics, even popular ones such as BLEU (Papineni et al., 2002[3]), are known to disagree with human assessment of translations to varying degrees and in different ways. Therefore, human assessment of translation quality has been established within machine translation as the most valid form of ground truth in tests (Callison-Burch, Osborne and Koehn, 2006[4]).

Third, and also of high importance for valid measurement, is the employment of suitable statistics that can accurately reflect the performance of systems. For example, a meaningful intuitive statistics for which established methods of statistical significance testing exist are superior to ad hoc measures. Examples include the mean or median for central tendency or Pearson correlation for association.

As a final consideration, since human evaluation of systems at scale takes substantial time and resources, many tests employ crowdsourced human assessments of system performance. This makes the validity of measurements highly dependent on strict quality checks. These test the reliability of human assessors for whom little to no verifiable information is known.

Direct assessment vs. crowdsourcing

The measure of choice for machine translation was coined by Conference on Machine Translation (Barrault et al., 2020[1]) as “direct assessment” (Graham and Liu, 2016[5]). It employs human assessors at scale with strict quality checks. Ratings of translation quality are collected on a 0-100 rating scale.

These ratings result in score distributions for systems for which accurate statistical tests can be applied. Such tests can avoid conclusions of differences in average ratings that are likely to occur simply by chance. Replication of experiments showed an almost perfect correlation with past results (Graham, Awad and Smeaton, 2018[6]).

Direct assessment has been used in machine translation for longitudinal evaluations that showed an average 10% improvement in system performance for machine translation of European languages over five years (Graham et al., 2014[7]). Furthermore, direct assessment made possible for the first time accurate comparison of human and machine translation system performance at the Conference on Machine Translation’s news translation task (Barrault et al., 2019[8]).

Results of tests showed that machine translation systems can outperform a human translator when sufficient training data are available. More recent results provide evidence that even professional human

translators can vary substantially in performance in tests. Consequently, a win over an individual human translator in a competition does not imply that a given system outperforms human translation or human translators in general (Barrault et al., 2020[1]).

Furthermore, direct assessment has been applied to additional AI tasks, such as video captioning (Awad et al., 2019[9]) and multilingual surface realisation (Mille et al., 2019[10]). Both of these Natural Language Generation research areas previously suffered from lack of reliable evaluation methodologies and low agreement in human assessment.

Avoiding evaluation pitfalls

Valid and reliable measures of system performance have not always been available within Natural Language Processing. This section provides examples of inaccurate or even misleading evaluation measures that were a necessary part of the process of producing more reliable measures. Several pitfalls, noted below, can be avoided when adapting evaluation techniques to new areas of AI:

- inappropriate application of statistical significance testing[4]
- application of statistics that allow machine learning algorithms to game the measure employed to evaluate systems[5]
- reporting results on selective subsets of data that show system/metric in most favourable light[6]
- lack of rigour in test settings that allow unfair advantage of systems due to unrealistic test data.[7]

Human-system hybrids and performance tests

In many AI applications, technologies will aid humans rather than replace them. As a result, pitting the performance of each against the other, only in isolation and working as entirely independent agents, could oversimplify reality. It would misjudge how technologies will be used in practice. Therefore, in addition to pitting human against machine in blind tests, performance of human/machine hybrids should ideally be measured where one or more human workers is *aided by* as opposed to being *replaced by* a given AI technology.

Relative participation of humans and machines

A hybrid test setting is more complex since it introduces the additional dimension of the relative participation of human and machine within completion of a given task. Participation can potentially range anywhere from *almost entirely automated* (with minimal human input) to *almost entirely manual* (with minimal AI).

Such a dimension brings the evaluation into a more realistic and therefore better setting where results can have a stronger impact. This added complication does raise questions. How should the degree of hybridisation be measured? Where would emerging AI and robotics technologies be placed along such a scale?

Creating a realistic, valid and reliable scale

A measure should make the resulting scale realistic, valid and reliable. The scale should provide a real-world reflection of how AI technologies rank against each other in terms of human participation. Valid measurements would accurately reflect the degree of hybridisation. The scale would also be highly reliable so that subsequent measurements would produce the same conclusions.

In terms of hybridisation, one method would be to measure the amount of human effort saved by a given AI technology. This could be estimated, for example, by giving sets of human workers tasks to complete with and without the aid of the AI technology. Tests could include measurements of, for example, average times saved for completion of a single task with the AI system. This could result in a time-saving scale for single-task completion for AI technologies (from no or little savings to vast savings).

Measuring the social value of AI technology

Measurement must also consider the scale at which society needs such tasks and the level of scaling of tasks possible. A task completed by AI could produce vast time savings but society or the workforce may not value this extra time. Indeed, a task that saves less time but is needed by many people or businesses or people worldwide will have more impact. Similarly, a technology with vast time-saving potential but is not scalable will also not have a high impact on employment or the workplace; measurements should try to reflect this.

An additional dimension is the potential social benefit of the AI technology. For example, a robot that can safely detect and deactivate bombs, aid a human perform life-saving surgery or prevent manipulation of the democratic process would rank highly on a scale of societal importance.

Measuring potential risks

Finally, an additional scale should ideally measure the risks of malicious applications of a given AI technology. Emerging Natural Language Processing technologies, for example, can generate fake news articles thought to be indistinguishable from human-authored news articles (Zellers et al., 2019[11]).

Such technologies might be widely deployable, scalable and operate almost independently of human input. However, identification of their malicious potential will increase the likelihood of developing preventative measures within society. This could aid against abuses such as manipulation of democratic processes. Numerous other examples of malicious AI exist such as deep fakes, drones and use of AI in the military.

In summary, tests can be devised to measure the human effort saved by assisting human workers with a given AI technology. There are four main considerations:

- time saved by hybridisation corresponding to a straightforward measurement of reduction in task completion time for a given task with and without the aid of the specific AI technology
- scalability, the need within society or the workforce for deployment of the technology at scale and whether scaling is possible
- importance of the task to individuals within society or society as a whole
- potential malicious impact of the AI technology and if this will require significant resources to deter or reduce risks on society or individuals within society.

Recommendations

As several of the earlier chapters have noted, computer scientists will argue that intelligence tests designed for humans are not suitable for measuring the performance of systems. This is primarily because human intelligence tests make assumptions about the basic abilities and skills of the test candidate; these are generally true of humans but not of AI. The main shortcomings with respect to natural language understanding are:

- Human intelligence tests require basic human abilities that AI systems do not have. For example, understanding questions in natural language and relating the meaning to real-world objects, and understanding how real-world objects might fit together, be manipulated and so on.

- Most areas of AI research do not attempt to obtain a general understanding of natural language. For example, they do not attempt to make the system complete tasks as a human would. Rather, they select a specific human intelligence task and work towards producing a system that can complete that single task.
- The vast majority of successful AI systems are evaluated in terms of completion of such tasks in isolation of other tasks. Humans are able to integrate separate basic skills, such as natural language understanding and object recognition. However, little AI research attempts to integrate multiple distinct task completion skills. For this reason, as well as other basic assumptions of human intelligence test results, these skills would be inappropriate for testing AI systems.
- An AI system trained to perform well on a specific human intelligence test is not a proof of intelligence in the way it would be of a human candidate. Even a minor change to the test format would likely lead to system failure due to its lack of general natural language understanding. Additionally, an AI system might perform well even on multiple human intelligence tests. However, without a function beyond these tests, it would be a system with few practical applications.

Highly sophisticated AI systems may one day possess basic human abilities such as general natural language understanding, general knowledge and understanding of the real world. They may learn to integrate such skills when faced with a new task. In this scenario, human intelligence tests would then be relevant.

However, measurement of system performance would need to mitigate against "gaming" the test. Just as humans can memorise answers to questions, a system can simply tune to the tests instead of demonstrating skill integration and general natural language understanding.

Although human intelligence tests are not appropriate for testing AI systems, other tests for humans could apply. Possibilities include vocational and educational tests, or indeed neuropsychological and developmental psychological tests that focus on the low-level skills possessed by most humans but not AI.

However, such tests are also likely to be misleading when directly applied to AI systems. Similar to human intelligence tests, such tests were devised with human candidates in mind. As a result, testing procedures include many assumptions about intelligence based on human intelligence alone. These assumptions, such as memory limitations and skills transfer, simply do not apply to AI.

Whether human or AI-specific tests are adapted or new tests are developed to evaluate AI systems, these should consider the following guidelines:

- **Avoid direct adoption of tests designed to test humans**

Assumptions about human candidates do not hold for AI systems.

- **Examine testing scenarios employed in each area**

These scenarios should include evaluation procedures in research papers, particularly for those employed in benchmark tasks.

- **Use human ratings of performance for evaluation**

This approach must ensure human assessors are blind to whether a system or other human is performing the task as opposed to metrics.

- **Employ realistic, unseen test data**

Freshly sourced data will help AI systems avoid tuning to the test data.

- **Include multiple humans in tests**

Humans should receive the same input data and/or environment as systems to represent human variance in performance, as well as to compare systems realistically.

- **Measure reliability of test results**

Reliability can be measured by repeating experiments with different human assessors or by measuring agreement of human assessors using inter-annotator and intra-annotator agreement measures, such as Kappa coefficient.

- **Repeat tests at regular intervals with new data and track improvements over time**
- **Report meaningful statistics and account for variance and statistical significance**
- **Measure performance of aid for a human rather than replacement**

Include a hybridisation scale that considers the performance of a given AI technology when aiding as opposed to entirely replacing a human.

- **Quantify the human effort involved to create key training data**

Include quantification of the human effort involved in creating training data for supervised AI technologies, monitoring and retraining systems.

- **Include scalability**

Technologies that can be deployed at scale are likely to have a higher impact on society.

- **Include both the potential and risk for society**

Include both the potential for positive impact and negative impact (e.g. malicious application of the AI technology).

References

Awad, G. et al. (2019), "Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval", *arXiv*, Vol. 2009.09984v1, http://dx.doi.org/arXiv:2009.09984v1. [9]

Barrault, L. et al. (2020), "Findings of the 2020 conference on machine translation", in *Proceedings of the Fifth Conference on Machine Translation*, Association for Computational Linguistics (online), https://aclanthology.org/2020.wmt-1.1. [1]

Barrault, L. et al. (2019), "Findings of the 2019 conference on machine translation", in *Proceedings of the Fourth Conference on Machine Translation*, Association for Computational Linguistics, Florence, http://dx.doi.org/10.18653/v1/W19-5301. [8]

Callison-Burch, C., M. Osborne and P. Koehn (2006), "Re-evaluatiing the role of Bleu in machine translation research", *Proceedings of the 11th Conference of the European Chapter*, Association for Computational Linguistics, Trento, https://aclanthology.org/E06-1032. [4]

Graham, Y. (2015), "Improving evaluation of machine translation quality estimation", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Association of Computational Linguistics, Beijing, http://dx.doi.org/10.3115/v1/P15-1174. [15]

Graham, Y. (2015), "Re-evaluating automatic summarization with bleu and 192 shades of rouge", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, http://dx.doi.org/10.18653/v1/D15-1013. [16]

Graham, Y., G. Awad and A. Smeaton (2018), "Evaluation of automatic video captioning using direct assessment", *PLOS ONE*, Vol. 13/9, pp. 1-20, https://doi.org/10.1371/journal.pone.0202789. [6]

Graham, Y. et al. (2014), "Is machine translation getting better over time?", *Proceedings of the 14th Conference of the European Chapter*, Association for Computational Linguistics, Gothenburg, http://dx.doi.org/10.3115/v1/E14-1047. [7]

Graham, Y. et al. (2020), "Assessing Human-Parity in Machine Translation on the Segment Level", *Findings of the Association for Computational Linguistics: EMNLP 2020*, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.375. [2]

Graham, Y., B. Haddow and P. Koehn (2020), "Statistical power and translationese in machine translation evaluation", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association of Computational Linguistics (online), http://dx.doi.org/10.18653/v1/2020.emnlp-main.6. [17]

Graham, Y. and Q. Liu (2016), "Achieving accurate conclusions in evaluation of automatic machine translation metrics", in *Proceedings of the 15th Annual Conference of the North American Chapter*, Association for Computational Linguistics: Human Language Technologies, San Diego, http://dx.doi.org/10.18653/v1/N16-1001. [5]

Graham, Y., N. Mathu and T. Baldwin (2014), "Randomized significance tests in machine translation", *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Baltimore, http://dx.doi.org/10.3115/v1/W14-3333. [13]

Koehn, P. and C. Monz (2006), "Manual and automatic evaluation of machine translation between European languages", in *Proceedings on the Workshop on Statistical Machine Translation*, Association for Computational Linguistics, New York, https://aclanthology.org/W06-3114. [12]

Ma, Q. et al. (2017), "Further investigation into reference bias in monolingual evaluation of machine translation", in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association of Computational Linguistics, Copgenhagen, http://dx.doi.org/10.18653/v1/D17-1262. [14]

Mille, S. et al. (2019), "The second multilingual surface realisation shared task (SR'19): Overview and evaluation results", in *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, Association of Computational Linguistics, Hong Kong, http://dx.doi.org/10.18653/v1/D19-6301. [10]

Papineni, K. et al. (2002), "BLEU: A method for automatic evaluation of machine translation", in *Proceedings of the 40th Annual Meeting*, Association for Computational Linguistics, Philadelphia, https://doi.org/10.3115/1073083.1073135. [3]

Zellers, R. et al. (2019), "Defending against neural fake news", in Wallach, H. et al. (eds.), *Advances in Neural Information Processing System*, Curran Associates, Inc., New York. [11]

Notes

[1] List adapted from http://https://2020.emnlp.org and intended only as a high-level summary of topics of interest in NLP as opposed to an exhaustive list. There is often overlap between different areas but no hierarchical relationship between tasks.

[2] A main venue for development of rigorous evaluation techniques was the Conference on Machine Translation (WMT) (Koehn and Monz, 2006[12]).

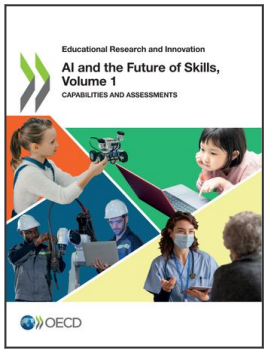
[3] It is not unusual for more than one system to be tied for first position in the competition.

[4] Competing statistical significance tests applied in machine translation evaluation were widely believed to provide substantially distinct conclusions. These were based on results of an oft-cited publication that claimed approximate randomisation to be superior to bootstrap resampling. Analysis in repeat experiments showed apparent differences in test results were simply due to an unwittingly bad comparison of one-sided and two-sided test results in experiments (Graham, Mathu and Baldwin, 2014[13]). Other analysis claimed that direct assessment introduced substantial bias in evaluation results. Yet this method had been employed since 2017 at the Conference for Machine Translation to produce official results. On further inspection and analysis of experiment data, claims of bias were revealed as unfounded due to the application of inappropriate analysis techniques (Ma et al., 2017[14]).

[5] An active area of Natural Language Processing is quality estimation. This involves application of machine learning algorithms to predict the quality of system-generated language, commonly applied to machine translation output. Within this area in benchmark tasks, systems were tested using measures such as mean absolute error. These yielded results in favour of systems that produced overly conservative quality estimates. Analysis of results showed an unfair advantage for systems that accurately predicted the mode of the test set score distribution and produced a conservative quality estimate located close to this mode. Subsequently, a more suitable measure was recommended that avoided the bias of previous results using a unit free measure, the Pearson correlation (Graham, 2015[15]). In evaluation of machine translation metrics, inappropriate statistical significance tests were applied to evaluation of metrics. This resulted in high proportions of over-estimates of statistically significant differences in performance, prior to identification and correction of this problem (Graham and Liu, 2016[5]).

[6] Inappropriate measures of Natural Language Processing were also widely applied in automatic summarisation. In this sub-discipline, an automatic metric was produced and widely adopted based on an apparent higher correlation with human assessment than BLEU. However, closer inspection of experiment data showed results were only presented for metric scores calculated on a subset of the relevant data. Consequently, they did not hold true for the full set of human annotations (Graham, 2015[16]).

[7] In machine translation research, an early Chinese to English machine translation system achieved performance on-par with a human translator. These results were identified as inaccurate for several reasons. First, there was a lack of context provided to human assessors in the evaluation. Second, it included reverse-created test data. Third, there was a lack of statistical power analysis to ensure sufficient sample size when concluding ties. Correction of such experiments is described in detail in Graham, Haddow and Koehn (2020[17]). They show that, contrary to initial claims, the human translator did not outperform the system when evaluated in the most appropriate way.



From:

AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:
https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Graham, Yvette (2021), "Assessing Natural Language Processing", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/fcd5e244-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at http://www.oecd.org/termsandconditions.