8 Towards a synthesis of language capability in humans and AI

Yvette Graham, Trinity College Dublin

Edited by: Nóra Révai, OECD

Language is a major part of human intelligence and researchers have been focusing strongly on developing such competences of machines. Natural Language Processing (NLP) technologies are a key area of artificial intelligence (AI). This chapter develops a conceptual framework of language competence that allows for comparing human and machine competences. It then maps major available language benchmarks on the framework and discusses the performance of state-of-the-art AI systems on a range of tasks in two broad domains: language understanding and language generation. This exercise is a first step in building an index of AI language capability.

Language ability in humans is a major part of intelligence. Throughout human history, and far earlier than the invention of the computer, people have fantasised about building robots that can communicate and understand language. The eventual success of computers to communicate flawlessly through natural language will greatly impact society and how people work. The field of Natural Language Processing (NLP) is concerned with allowing computers to process and simulate an understanding of natural language in spoken or written form. NLP thus forms a major component of artificial intelligence (AI), but itself comprises several different sub-areas that separate NLP into problems or *tasks*. Each of these areas relates to some notion of human language competence. This raises the question: how can human language competence (which itself varies from one individual to another) be compared with state-of-the art performance of NLP systems?

This chapter compares human levels of competence with NLP system performance levels using benchmarks from the field of computer science. In terms of the range of human language competence levels, the analysis concerns the mainstream working population and education of the general future workforce (i.e. it reflects the competences of any human who does not possess a severe disability).

Benchmark tasks: Narrow versus strong AI

Each research area in NLP includes one or more *benchmarks* or *shared tasks* that aim to compare the performance of competing approaches to determine which methods have most promise. A task is defined with a core research goal of automating some form of language processing in a specific way and with respect to a specific domain of language (e.g. translating news documents from German to English). To correctly interpret NLP benchmark test results, *narrow* AI that focuses on solving individual tasks must be distinguished from *strong* AI that aims to simulate *general purpose intelligence*. Technologies are developed and tested in isolation from other tasks so almost all NLP benchmarks currently form part of narrow AI. This means that NLP systems can legitimately be tested on their performance of a single language task within a single domain (e.g. news text, medical documents, literature or scientific papers). The performance achieved – even if very high – is limited to the evaluation setting, which is restricted to that specific domain and task. Working on individual problems in NLP in isolation from other tasks allows for progress, making an insurmountable mountain climbable through mostly independent routes.

While it is important to interpret NLP success within the context of narrow AI, components of a successful NLP system can often be applied to a new task, domain or language. Such components may very well form part of an eventual general purpose language AI. How the research community can ever achieve a general purpose (or strong) language AI is still a question. The key to this may lie in the underlying technologies that have proven successful (or will be) across multiple NLP tasks, languages and domains (see recent development in Box 8.1).

Box 8.1. Transformer models

The recent development of neural NLP architectures as transformer models has helped the move towards general purpose AI. The emergence of this paradigm has been a game changer in terms of NLP system performance, resulting in discussions around human parity at several tasks. A transformer model learns to understand and represent the meaning of language from an exceptionally large volume of raw text with no human annotation.

The most well-known such model developed by OpenAI, GPT-3 (and its first publicly released version, ChatGPT) is trained on half a trillion words (or tokens) of English, sourced from a combination of webpage content and books. This massive language model requires training only a single time and can then be applied repeatedly and in a wide range of distinct tasks. The model can be deployed to successfully automate a range of distinct language tasks, such as Machine Translation, Named Entity Recognition, Question Answering and Speech Recognition, among others.

Typically, a single pretrained language model can be used with further fine-tuning, which requires only a relatively small data set. This model is referred to as a *transformer* since it transforms the information it has learnt from the pretrained model to a more specific task. The technology behind this approach is based on the structure of the human brain in the form of neural networks.

Conceptual framework of language competences

Comparing human language competence to NLP benchmarks requires bridging the gap between the general understanding of description and analysis of human language competence and NLP research.

Annex Table 8.A.1 presents a list of main NLP research areas with the equivalent human skill each task aims to automate. Human language competence is generally not analysed by tasks. Rather, it is usually divided into the four competence areas of reading (HR), writing (HW), listening (HL) and speaking (HS). Language competence is then described on a range from low level of competence in a given native language (e.g. with no reading or writing ability) to high proficiency in all four categories in the native (or another foreign) language.

To map human language competences to NLP areas, two high-level groups can be formed:

- *Language understanding*: NLP tasks that correspond closely to reading or listening competence in humans (understanding/comprehension tasks); both require interpretation of *input.*
- *Language generation*: NLP tasks that correspond to writing or speaking require the system to *output* language.

Some NLP tasks correspond to a combination of language understanding and generation. The distinction between reading or listening and writing or speaking is only a matter of input/output formats moving from text to speech. The core technology or research problem, such as translation, largely remains the same regardless of input or output format.

However, the format that an NLP system receives or produces can impact system performance. A move from text input/output to spoken input/output usually involves a *decrease* in system performance, as spoken input is less predictable and more difficult to process than textual input. The opposite is true for humans: understanding or generating spoken language instead of reading or producing text is usually easier because it does not require competence in literacy. Figure 8.1 shows this relationship. There are exceptions to this general rule. For example, machine translation and interpreting are more challenging for both system and human translation in spoken form.

Figure 8.1. General relationship between human language and NLP difficulty levels with respect to the input and output format moving from text to speech and vice versa

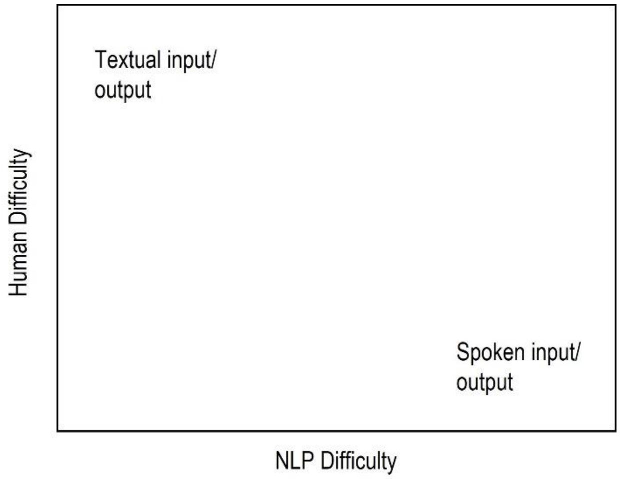


Table 8.1 provides a rough guide of how core NLP research areas relate to the type and competence level required of a human to perform the equivalent task. Some tasks require different abilities from systems and humans. For example, Speech Recognition for humans corresponds to the task of understanding spoken language and, as such, generally requires low level of language ability. However, NLP systems are tested by the production of a transcription, i.e. a written output.

Table 8.1. NLP research areas with type and level of language competence required for humans

Minimum Human Competence Level	Reading/Listening	Both Reading/Listening and Writing/Speaking
Below Average	Emotion-cause Pair Extraction Event Extraction Humour Detection Reasoning (basic) Visual QA	Dialogue (open domain) Speech Recognition
Average-High	Anaphora Resolution Natural Language Inference Part of Speech Tagging Reasoning (advanced) Sentiment Analysis Text Classification Topic Modelling	Explanation Generation Grammar Correction Keyword Extraction Lexical Normalisation Punctuation Restoration Question Answering Reading Comprehension Sentence Compression Summarisation Text Diacritisation
Specialist	Authorship Verification Automated Essay Scoring Information Retrieval Semantic Parsing Syntactic Parsing Relation Extraction	Dialogue (task-oriented) Machine Translation Text Style Transfer

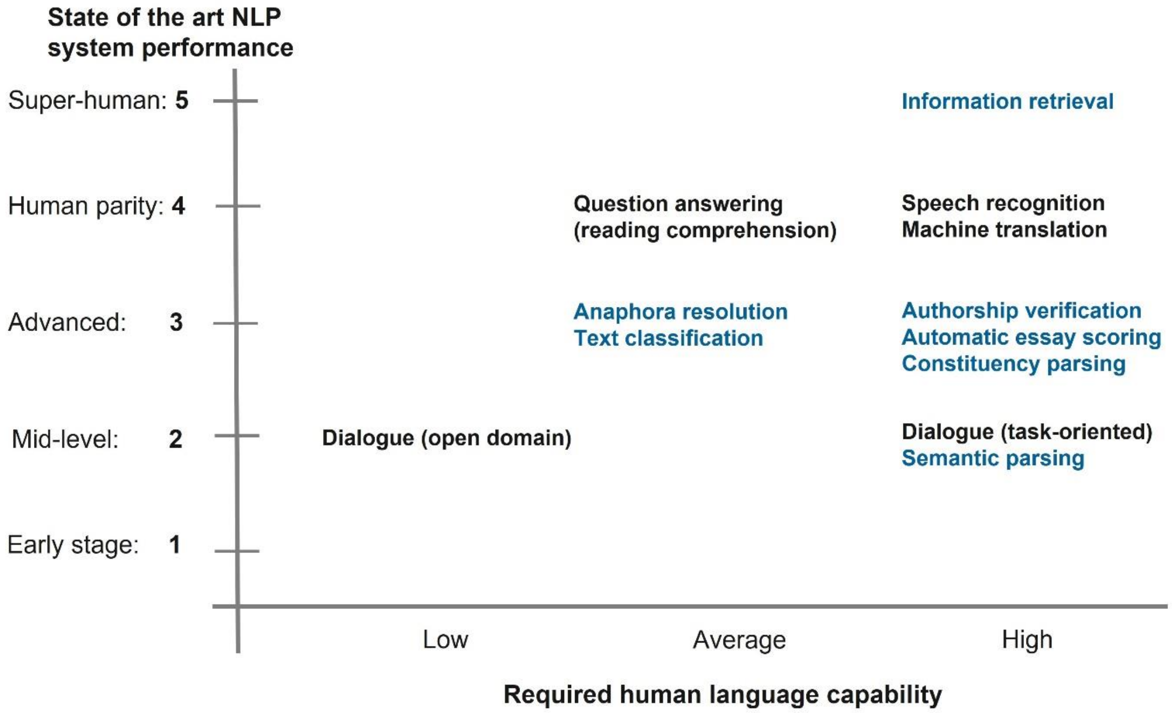
Mapping major language benchmarks to the human language competence framework

This section maps the performance level of state-of-the-art systems in each research area to human competence levels required for a corresponding task. Figure 8.2 comprises three basic performance levels corresponding to: i) *early-stage research;* ii) *mid-level* research; and iii) to *advanced* research. There are two additional categories: iv) tasks in which *human parity* discussions have begun and v) areas with consensus that systems have already achieved *super-human* performance. The figure reflects state-of-the art research in 2021; the performance of systems is likely to improve over time.

Some NLP tasks, such as Machine Translation (MT) or Speech Recognition, are application-driven, i.e. they correspond to human tasks. For these, it is natural to ask: *how does state-of-the-art AI compare to the ability of a human completing the same task*? With these tasks, researchers aim to reach performance equivalent to that of humans (human parity), or even surpassing the capability of all humans (super-human performance).

However, not all NLP tasks are driven by direct application. Some tasks aim to automate an annotation process that usually requires a human to complete. The ultimate aim is simply to annotate/label data correctly as a human would have. The human annotator in this case provides the gold standard annotations against which the outputs of the NLP technology are judged (as either correct or incorrect). For such tasks, questions around human-parity or super-human performance are arguably not highly relevant. Therefore, these research areas are only classified in the first three categories.

Figure 8.2. Relationship between required minimum human language competence level and state-of-the-art NLP system performance for a sample of NLP tasks



Note: This is a rough guide based on state-of-the-art NLP research in July 2022. Blue benchmarks refer to language understanding, black ones to language generation.

Figure 8.2 provides a rough guide of the current relationship between NLP state-of-the-art performance and the minimum human language competence required to complete that task, with reference to a sample of NLP tasks. The sections that follow describe each task shown in the graph and explain why it was placed at the given performance level based on recent NLP benchmarks. A discussion of language understanding is followed by language generation tasks.

Language understanding: AI vs. human

Overall, Figure 8.2 suggests that AI systems perform better in language understanding tasks than a human with low level reading and listening competence. Many research areas are in an advanced stage, with Information Retrieval (IR) being at the top – super-human – level.

Understanding words in their context

Task definition: Anaphora and Coreference resolution

The meaning of many words cannot be interpreted correctly isolated from their context. For example, *it* commonly refers to a noun mentioned earlier in the text. To translate the word *it* from English to German, the gender of the word *it* refers to in the text must be known. For example, in "*The car was brand new, but he still allowed me to drive it*", the gender of *car* in German must be known. Understanding this link between words in a text/speech is relatively easy for humans but can be challenging for NLP systems.

Anaphoric words are expressions like *it* in the example above, whose interpretation depends on the context. Anaphora and coreference resolution are NLP tasks that aim to identify entities referred to within text or spoken language by such anaphoric words or expressions.

System performance on benchmarks

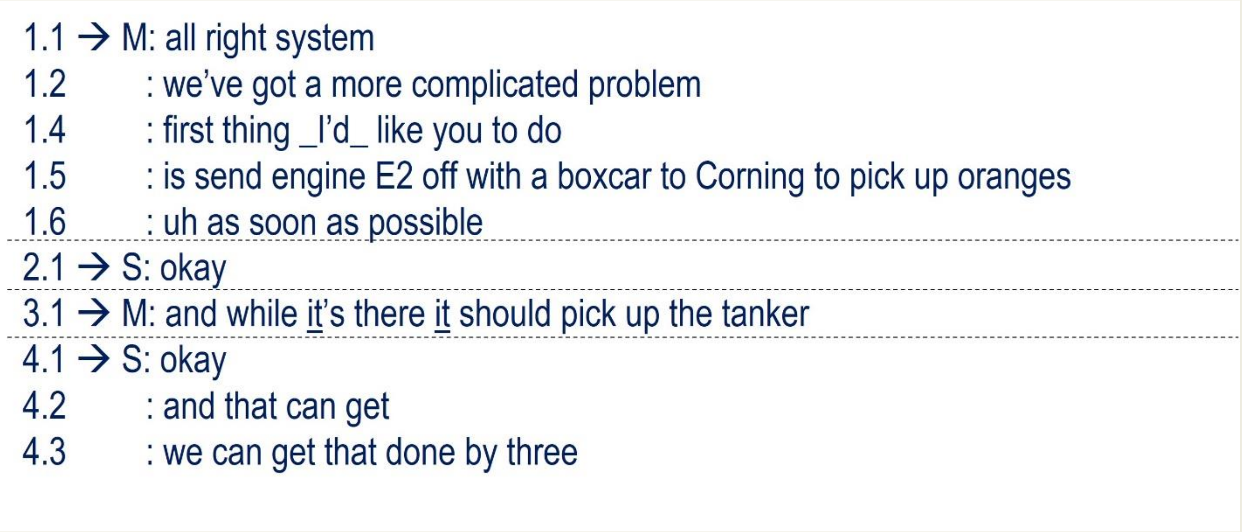
The domain of anaphora and coreference resolution includes an extensive range of entities and numerous benchmarks that test system performance (see Sukthanker et al. (2020[1])). A main dataset is the Ontonotes corpus (Weischedel et al., 2013[2]). This was created to develop methods of automatic coreference in order to link all the specific mentions in a text that refer to the same entity or event. In addition, the corpus was annotated to distinguish between different types of coreference. Texts automatically annotated in this way are likely to help other NLP tasks learn to correctly process multiple mentions of the same entity.

State-of-the-art systems based on a transformer architecture achieve high performance with respect to Ontonotes: an F-score (a score that combines precision and sensitivity) of approximately 81% (Dobrovolskii, 2021[3]). Despite a lack of human performance estimates for the task, performance can still be gauged to some degree. The highest performing systems surpass the performance of someone with low reading and listening competence. However, they are unlikely to surpass that of a human with an average or high language competence.

Box 8.2. Example of anaphora and coreference resolution

In the example dialogue below taken from the TRAINS corpus (Poesio et al., 2016[4]), the personal pronoun *it* refers to two distinct objects in utterances 3.1 and 5.4.

Figure 8.3. Example dialogue from the TRAINS corpus



1.1 → M: all right system
1.2 : we've got a more complicated problem
1.4 : first thing _I'd_ like you to do
1.5 : is send engine E2 off with a boxcar to Corning to pick up oranges
1.6 : uh as soon as possible

2.1 → S: okay

3.1 → M: and while it's there it should pick up the tanker

4.1 → S: okay
4.2 : and that can get
4.3 : we can get that done by three

Source: Adapted from Poesio et al. (2016[4]): *Anaphora Resolution, Springer*.

Structuring and organising text into categories

Task Definition: Text Classification

Text classification systems aim to automatically categorise a given sentence or document into an appropriate category. They can help organise and structure any kind of text, such as sentences, documents and files. The number and types of categories depend on the application and associated dataset. For example, news articles can be categorised by topic, sentences can be labelled with the emotions expressed by them or as grammatical or ungrammatical. Systems can then be trained on such datasets to classify unseen sentences in these categories. As such, text classification overlaps with other more specific NLP research areas, such as sentiment analysis and grammar correction.

System performance on benchmarks

Text classification is a long-established area of NLP with extensive work and progress likely due to the wide availability of substantial data for training and testing systems. One of the most widely used set of datasets is the General Language Understanding Evaluation (GLUE) (see Box 8.3).

Box 8.3. GLUE benchmark

The General Language Understanding Evaluation (GLUE) benchmark comprises nine datasets for classification of sentence of English. The name of each dataset and the task associated in Table 8.2 is adapted from Wang et al. (2018[5]).

Table 8.2. Datasets in the GLUE benchmark

Name	Description
CoLA (Corpus of Linguistic Acceptability)	Determine if a sentence is grammatically correct or not.
MNLI (Multi-Genre Natural Language Inference)	Determine if a sentence entails, contradicts or is unrelated to a given hypothesis.
MRPC (Microsoft Research Paraphrase Corpus)	Determine if two sentences are paraphrases from one another or not.
QNLI Question-Answering Natural Language Inference)	Determine if the answer to a question is in the second sentence or not.
QQP (Quora Question Pairs2)	Determine if two questions are semantically equivalent or not.
RTE (Recognising Textual Entailment)	Determine if a sentence entails a given hypothesis or not.
SST-2 (Stanford Sentiment Treebank)	Determine if the sentence has a positive or negative sentiment.
STS-B (Semantic Textual Similarity Benchmark)	Determine the similarity of two sentences with a score from 1 to 5.
WNLI (Winograd Natural Language Inference)	Determine if a sentence with an anonymous pronoun and a sentence with this pronoun replaced are entailed or not.

Source: Wang et al. (2018[6]) : *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. https://doi.org/10.18653/v1/w18-5446.

State-of-the-art results can vary depending on datasets but range from approximately 75% to 90% in terms of accuracy. Despite high performance, discussions of human parity at this task have not yet taken place. This is likely because the task is closer to an annotation (labelling) task than an application task. In other words, human annotations for this task are deemed correct, leaving aside disagreement between annotators that can occur. This area can be classified at an *advanced* performance level, with at least average human language competence required to perform the same task.

Understanding the meaning and role of expressions

Task Definition: Semantic Parsing

Understanding the meaning of expressions within and across sentences can be the basis for translation, answering questions and reasoning. Semantic parsing aims to automatically annotate sentences of a given natural language with *formal meaning representations*. A typical example is to automatically identify the subject, object and indirect object in a sentence of English. For example, the sentences "*Mary gave John the letter"* and "*Mary gave the letter to John"* are identical in terms of semantic structure, despite the order of words being different.

System performance on benchmarks

Semantic parsing is an easy task for humans, even someone with low literacy skills can figure out who did what to whom in the sentence above. However, the problem is more challenging for machines. They require parsing the sentence with the grammar of the specific language to determine this simple information from a simple sentence. There are numerous possible formalisms for semantic parsing of natural language, and studying a single natural language, such as English, only illustrates a fraction of the features of language in general. Further challenges include long-distance dependencies between words (i.e. links between words that are far from each other in the sentence) and languages that suffer from data sparseness due to rich morphology, such as Arabic, Czech and Turkish. Interestingly, the original applications of semantic parsing, such as MT, have had much more success without the integration of semantic representations.

Parsing language has received a good amount of attention over the years within the NLP community. A number of datasets for training and testing semantic parsers exist for a range of different formalisms. Propbank (Palmer, Gildea and Kingsbury, 2005[7]) is a corpus annotated with predicate argument structure. For its part, Framenet is a major dataset (Baker, Fillmore and Lowe, 1998[8]) in which the usual unit of meaning – a word – is replaced with other lexical units and frames. More recently, a project known as Universal Dependencies has made gallant strides towards developing a cross-linguistically consistent treebank annotated with semantic roles with the goal of multilingual parser development (de Marneffe et al., 2021[9]).

Although extensive time and energy have been invested in this research area and high accuracy achieved in shared tasks, much work to date has focused on repeated tests on the same dataset. In addition, tasks have overly focused on English, a language that probably poses far fewer challenges than morphologically rich languages. As a result, correctly classifying semantic parsing technologies is difficult. This area could be best placed as having *mid-level* performance to consider the remaining challenges for developing technologies to a large number of languages. Corresponding minimum human level of language competence is high (specialist) in terms of formal annotation (and not simply understanding a sentence).

Understanding sentence structure

Task Definition: Constituency Parsing

Analysing a sentence by breaking it down into sub-phrases of different grammatical categories (e.g. noun phrases, verb phrases) can help in more complex language tasks, such as grammar checking, semantic analysis and Question Answering (QA) (Jurafsky and Martin, 2023[10]). Constituency parsing is the task of automatically annotating sentences of natural language with a phrase structure grammar. Figure 8.4 shows a constituency parse tree corresponding to a phrase structure grammar of an example sentence.

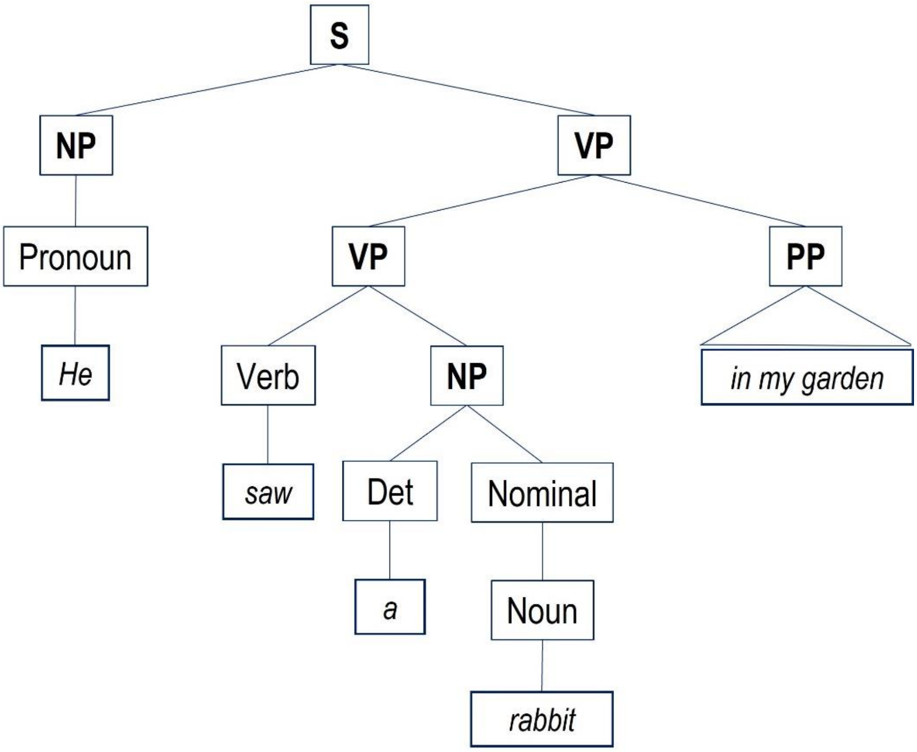
System performance on benchmarks

Two major datasets for testing constituency-parsing technologies include the English Penn Treebank and the Chinese Treebank, with highest performing systems achieving approximately 97% and 93%, respectively (Mrini et al., 2020[11]).

Since constituency parsing is essentially automated annotation of data (annotation task), the question of human parity is unlikely to be discussed. The question is not whether systems have reached human parity in constituency parsing itself. It is more interesting to ask whether systems have helped reach human parity with respect to a larger application task.

This research area falls into *advanced* performance level with respect to parsing English text, since accuracy is exceptionally high. However, this research area has shown an excessive focus on parsing results for English.

Figure 8.4. Sample constituency parse tree of English sentence



Source: Adapted from Jurafsky and Martin (2023[10]): *Speech and Language Processing, Pearson Education Inc.*

Assessing student essays

Task Definition: Automatic Essay Scoring

Automatic Essay Scoring (AES) is the task of automatically assessing student essays and has applications within education. For example, AES has great potential to allow students to get feedback about the quality of their writing without requiring teaching professionals' time and resources.

System performance on benchmarks

The Automated Student Assessment Prize dataset, released by Kaggle (www.kaggle.com), is a main resource for training and testing AES systems (www.kaggle.com/competitions/asap-aes/data). The dataset contains eight essay sets, with essays ranging from 150 to 550 words per response, written by students in US grade levels from grades 7 to 10. Essays within the dataset are hand-graded and double-scored to test rater reliability.

Results for state-of-the-art systems show performance at approximately 80% weighted kappa. Discussions of human parity in this area are unlikely to develop. Arguably, they are not highly relevant due to the nature of automating a task for which human scores provided by a qualified teacher are considered valid and correct. AES can be considered as having *advanced* system performance while requiring high (specialist) human language competence.

Identifying the author of a document

Task Definition: Authorship Verification

Authorship verification (AV) is an NLP task to automatically determine if a new unseen document was authored by an individual already known to the system. The task has applications in detecting plagiarism and in data analytics for commercial systems that aim to profile users based on the content they have authored on line, among others.

System performance on benchmarks

Despite the huge potential of NLP technologies to successfully categorise textual content according to author, testing in this area has been limited by lack of data. Data needed for this task would ideally comprise text authored by a large set of individuals (*n*), each of whom had authored a large set of documents (*m*), yielding a potential dataset containing *n x m* texts to train systems. Such datasets are unfortunately not readily available. Furthermore, even given the availability of such data for a single domain, systems should be tested on texts/documents across a range of domains of language to verify that results achieved for one domain can be achieved in another.

To work around the lack of ideal training and test data, benchmarks have taken the available data and simplified the task. They only require systems to determine if the same individual authored two given documents (Göeau et al., 2021[12]). Data for benchmarks are taken from the fanfiction domain. Fanfiction refers to new stories authored by fans of a well-known show/book that include its characters (Kustritz, 2015[13]). Since these kinds of data provide multiple stories authored by the same individual, fanfiction lends itself to training and testing AV systems.

Despite systems generally only being tested within the fanfiction domain, there is no reason to believe that systems would not achieve similar results in other domains provided that data are available. However, measures applied in benchmarks for this task are not straightforward to interpret, nor do they readily map to human competence. For example, systems are permitted (and somewhat encouraged) to sit on the fence for test items and submit decisions of 0.5 probability to indicate indecision about a difficult case. This results in metrics reported on distinct numbers of outputs. This, in turn, gives an advantage to systems that sit on the fence more often, making comparisons across systems difficult. Consequently, it is difficult to gain a simple intuition about how often systems correctly identify authors of documents.

In addition, no attempts have been made to estimate the degree to which humans can determine if two stories had the same author. This leaves the question about the performance level of state-of-the-art AV systems. The organisers of the latest shared task report that the F-score of the best system (about 95%) is highly encouraging. However, they note that these results may not hold for other domains and the test domain may simply be too easy.

This task was placed as having *advanced* system performance, while keeping in mind the above-discussed caveats. Even humans with high language competence might find this task exceptionally challenging. If this is the case, systems could perform better than humans at AV.

Finding material relevant to a question

Task definition: Information Retrieval

Information Retrieval (IR) is defined as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) (Manning, Raghavan and Schütze, 2008[14]). IR is not generally considered to be a sub-discipline of NLP, as the methods proven successful especially in the early days have had relatively little in common with

NLP systems. IR has focused on word counts in documents (or statistics), compression techniques and efficient algorithms to speedily sift through enormous quantities of data to respond to the information need of a user.

System performance compared to humans

Even the most basic IR algorithms from the early days have already surpassed human limits for IR due to the scale of data needed to be searched. Benchmarks are thus less relevant to understand the extent to which IR has achieved performance comparable with humans. IR systems clearly have super-human performance. A human carrying out the IR task corresponds to the task traditionally carried out by a trained librarian, and thus requires high (specialist) language competence.

Language generation: AI vs. Human

The competence to generate language whether in spoken or written form generally requires language understanding in real life and workplace language tasks. Figure 8.2 suggests that AI performance in language generation ranges from mid-level to human parity depending on the research area.

Dialogue in an open domain

Task Definition: Open-Domain Dialogue

Dialogue systems aim to automate the art of human conversation. Dialogue research is generally split into two distinct areas. Open-domain dialogue automates conversation about any topic of interest, also referred to as chit-chat models. Task-oriented dialogue completes specific tasks through automated conversation (see section Dialogue in a narrow domain).

System performance on benchmarks

A main venue for evaluating dialogue systems is the Conversational Intelligence Challenge (Convai) (Burtsev et al., 2018[15]; Dinan et al., 2019[16]). Evaluating open-domain dialogue systems is challenging because it requires human evaluators to talk to chatbots about an open or prescribed topic before rating their quality. In fact, human evaluations in the original competition were found to be unreliable, and the results discussed here use an alternate source for evaluation results.

A recent evaluation of state-of-the-art models, which was replicated to ensure reliability, showed that the highest performing models rated by human judges perform at approximately 52% (Ji et al., 2022[17]).[1] State-of-the-art models in this area are based on transformers that learn from extremely large language models. They produce highly fluent output that often is appropriate for the input provided by its human conversation partner. However, despite fluent output, models still lack consistency in conversations and the ability to reliably incorporate knowledge or learn new information from conversations. We place this research area into the category of *mid-level* performance in relation to other NLP tasks. This task needs low language competence from humans.

Understanding and transcribing spoken language

Task Definition: Speech Recognition

Speech recognition systems aim to take in a speech signal from one or more speakers and produce a textual transcription of the language that was spoken. Much of the research developed in speech recognition has been applied successfully to other NLP tasks. As mentioned previously, systems are tested on their transcription ability, while testing humans' speech recognition does not require written literacy. This complicates the comparison, as for machines it involves the generation of language output.

System performance on benchmarks

Speech recognition has received a wealth of attention over the years, partly due to the availability of data for training and testing statistical systems. There are a large number of benchmarks and datasets for this area and it is a leading area of NLP in terms of system performance. Indeed, many other research areas adapt methods first successful in speech recognition. Researchers are discussing human-parity and even super-human performance of systems, but challenges nonetheless remain. It is yet to be shown that the accuracy of developed speech recognition technologies exceeds that of a human transcription for all kinds of spoken language data. Speech recognition was placed at the human-parity level. Since almost all humans (without serious disabilities) have this capability, it only requires a low human language ability.

Answering questions based on reading comprehension

Task Definition: Question Answering

Question answering (QA) is the task of automatically finding answers to questions or identifying when a question cannot be answered due to ambiguity or lack of information to provide an appropriate answer. QA overlaps with other NLP research areas such as reading comprehension. Rather than mere IR from a pre-engineered knowledge base with facts, QA with reading comprehension means a system can comprehend a text and absorb the knowledge without prior human curation.

System performance on benchmarks

QA is an extensively studied area of NLP. There is a vast number of datasets and benchmarks for evaluating systems, likely exceeding 100 distinct datasets for testing some form of QA system. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016[18]), for example, is a reading comprehension data set containing a sample of questions and answers about Wikipedia articles. In SQuAD, systems must either find the answer to each question in the corresponding article or identify that the question is, in fact, unanswerable. Results for state-of-the-art QA systems for SQuAD and other QA datasets such as CommonsenseQA reveal excellent system performance and have led to discussions of systems reaching human parity. The human language competence required for effective QA is average to high.

Dialogue in a narrow domain

Task Definition: Task-oriented Dialogue

Certain conversations pertaining to a specific task, such as asking about the weather and giving navigation instructions to a driver, happen regularly in many people's lives and are thus worth automating. Task-oriented dialogue systems aim to help users complete a task of some description through automated conversation. This can be in the form of actual spoken interaction with the system or a text-based interface or task-oriented chatbot.

System performance on benchmarks

Given the large number of use cases, testing for task-oriented dialogue systems is still limited by available training and test datasets. A main dataset for task-oriented dialogue is Key-Value Retrieval Networks (KVRET), which contains more than 3 000 dialogues across the in-car assistant domain, calendar scheduling, weather IR and point of interest navigation in English (Eric et al., 2017[19]). Models are evaluated using entity-F1, a metric that evaluates the model's ability to generate relevant entities from an underlying knowledge base and to capture the semantics (Eric et al., 2017[19]). The current best task-oriented dialogue system achieves an entity F1 score of approximately 71% (Xie et al., 2022[20]).

Humans have achieved 75% on such tasks, which suggests impressive performance of state-of-the-art systems. However, AI systems were evaluated on datasets similar to training data, instead of new unseen test data. Given the limitations of evaluations and available datasets system performance must be judged with caution. Task-oriented dialogue is still extremely challenging, corresponding most closely to *mid-level* performance. On the human side, such tasks often require a specialist and thus correspond to high-level language competence.

Translating text

Task Definition: Machine Translation

Machine Translation (MT), i.e. automatically transferring the meaning of text or speech from one natural language into another, is one of the earliest AI language tasks. It presents many challenges. First, there are, in theory, an infinite number of possible input sentences. This means that translating sentences requires breaking them down into short units to find a *phrase* that has been seen in the training data (or is in the database), and translate it by components (slicing and dicing). Second, there are many ways to slice and dice a single sentence, and many possible ways to translate each of those slices. This results in a vast number of possible outputs for every input sentence. Determining which of these is the best translation is a main challenge in MT.

While phrase-based MT has been a highly successful and long-standing approach to translation (Koehn, Och and Marcu, 2003[21]), transformer models have recently made substantial advances. Despite this progress, challenges remain. These include sentences containing long-distance dependencies between words, issues relating to incorrectly translated pronouns (discussed in the section on anaphora resolution), the translation of languages with rich morphology and languages with low amounts of training data.

System performance on benchmarks

Generally in MT, the two languages between which a system translates is known as its *language pair*. Besides the many challenges that lie within the task of MT, language pairs create another problem with respect to reporting system performance. MT performance can be easily gauged from benchmark results for a language pair that, for example, translates from German (de) to English (en). However, assessing translation performance between *any* two natural languages is less straightforward. This is partly because of the large number of possible language pairs, but more importantly because of big performance gaps between language pairs. MT has excellent performance for some pairs, for which large datasets are available and researchers have worked on them extensively. Conversely, performance is much lower for pairs that have no data or systems for testing.

The data-driven methods the MT research community has most focused on are language pair independent. These state-of-the-art methods learn to translate from large corpora (as opposed to hand-crafting large sets of rules). This means that once training data for many language pairs become available, a high performing system can be built relatively quickly using already developed code. Therefore, with some degree of caution, results can be extrapolated on state-of-the-art methods for which training data already exist for any language pair.

Recent advances have resulted in discussions about human parity in many leading benchmarks. The news translation task at the Conference on Machine Translation (WMT) (Akhbardeh et al., 2021[22]), for example, has highly valid and reliable test results. This is because substantial effort is put in developing new test data before each annual competition, ensuring that the test data are truly unseen. In addition, the task employs human evaluation as opposed to automatic scoring, with both systems and human translators included in competitions in a blind test. WMT benchmarks have shown that on average the best system(s) achieve performance on-par with a human translator.

Update of AI language competences post ChatGPT release

The first draft of this chapter was written in summer 2022. In November 2022, ChatGPT was released, marking a significant milestone in AI. By January 2023, it garnered over 28 million daily visits, which many describe as the highest impact AI technology advancement. OpenAI presents ChatGPT as a step towards Artificial General Intelligence. It can hold high-quality conversations, answer follow-up questions, admit mistakes and challenge incorrect assumptions. Many users find its simulated intelligence convincing, able to answer any question with higher fluency and linguistic ability than many speakers of English. ChatGPT has been tried for tasks like writing job applications, emails and even academic essays.

Despite the apparent advancement in recent dialogue systems, such as ChatGPT, **it is not possible to track significant improvement in AI’s language capabilities over one year**. This is because new models have not yet been independently tested on language benchmark tasks.

This brief section re-evaluates the benchmarks used to measure AI language capabilities and systems’ performance on the old and new benchmarks in light of the advancements.

Evolution of benchmarks

While all benchmarks discussed above remain pertinent, some already existing and new benchmarks have gained importance. SuperGlue, an advanced version of the original GLUE benchmark, has emerged as particularly significant. It offers a comprehensive metric for gauging progress toward general-purpose language understanding for English (Wang et al., 2018[5]). SuperGlue's new tasks provide a better measure of AI’s underlying capabilities compared to its predecessor. Tasks included in SuperGlue are challenging for AI but solvable by most college-educated English speakers. SuperGlue features:

- more challenging tasks, retaining the two toughest from GLUE and adding others based on their difficulty for current NLP systems
- diverse task formats, expanding from just sentence pair classification to coreference resolution and QA formats
- comprehensive human baselines
- enhanced code support
- refined rules to ensure fair competition.

Importantly, SuperGlue provides human performance estimates. In 2019, the average accuracy across eight tasks was 71.5, compared to a human score of 89.8, indicating an almost 18 percentage point difference (Wang et al., 2018[5]). More recently, in October 2023, the leaderboard shows significantly improved performance figures for more competitive systems, reflecting a large increase in performance since 2019. However, numbers need to be interpreted with a degree of caution as submissions to SuperGlue are permitted up to 6 times per month, so at least some degree of tuning to the test could have taken place over the past 4 years.

However, SuperGlue has some limitations. First, tasks that require domain-specific knowledge were not included. Thus, SuperGlue is not able to assess AI’s capability to integrate knowledge and language competences. Yet, this is necessary to closely mimic human intelligence. Second, some human evaluation estimates rely on anonymous crowd-sourced data, which might be of lower quality than expert annotation, potentially affecting human performance estimates.

Other new benchmarks aiming to test general-purpose intelligence have also emerged this past year. One that has gained significant attention is the Beyond the Imitation Game Benchmark (BIG-bench) (Srivastava et al., 2022[23]). BIG-bench is a collaborative benchmark with more than 200 tasks, intended to probe large language models and extrapolate their future capabilities. However, it moves beyond language processing tasks and thus is not considered in this update. Similarly, benchmarks that focus on evaluating language models in a zero-shot setting (e.g. LMentry) were not considered either. These benchmarks test a pretrained language model's general understanding of language through transfer learning as opposed to their performance on a specific task they were trained on.

To date, there are no appropriate benchmarks available to assess technologies like ChatGPT as a single system that covers a broad spectrum of NLP tasks. Future research would benefit from benchmarks designed for general-purpose language AI systems, enabling more accurate comparisons between ChatGPT, its competitors and human capabilities.

Evolution of system performance in language

A system that aims to be an Artificial General Intelligence is not designed for one specific NLP task only. Rather, it aims to simulate human language ability across various tasks. ChatGPT focuses on a subset of NLP tasks: question answering, summarisation and general language generation, including letter writing, report writing and storytelling. Importantly, it aims to master dialogue in an open domain. Its success stems from its ability to integrate multiple tasks seamlessly, offering a user-friendly interface. However, when comparing ChatGPT to individual NLP benchmarks and human language capability, it is essential to assess its performance in specific tasks.

Overall, almost all NLP tasks explored in this paper remain at the same level of performance after about a year of development. There are two tasks that need revisiting because of the breakthrough with large language models: Dialogue in a narrow domain (task-oriented dialogue) and in an open domain.

Regarding task-oriented dialogue, AI performance has improved on a limited number of tasks that can be performed to a high standard through conversational instruction by systems such as ChatGPT. However, systems still cannot perform successfully on a wide range of tasks through dialogue with users. Thus, overall performance remains at mid-level.

As a fluent conversational agent, ChatGPT and other publicly available dialogue systems based on transformers and large language models constitute a breakthrough. However, ChatGPT still faces challenges in numerous tasks, like applying knowledge learnt in conversations with users, resolving references and integrating real-time world knowledge. Thus, overall performance of AI in open-domain dialogue may not yet have moved to an advanced level.

Other breakthrough technologies beyond ChatGPT

Following the release of ChatGPT, OpenAI introduced GPT-4, a significant advancement in NLP. GPT-4 processes both image and text inputs to produce textual responses. While independent evaluations of GPT-4 are still pending, OpenAI has shared results that suggest advanced performance. GPT-4 is reported to achieve human parity on various professional and academic benchmarks in terms of factuality, steerability and ethically declining certain queries.

GPT-4 is said to pass simulated bar exams, scoring within the top 10%, a significant improvement from its predecessor GPT-3.5, which scored around the bottom 10% (OpenAI, 2023[24]). However, it's essential to note that these are simulated tests, suggesting they might differ from real-world exam settings. In addition, the results discussed here are OpenAI's in-house tests that have not yet been independently verified and only limited details are publicly available. Despite these accomplishments, OpenAI acknowledges GPT-4's limitations, especially when compared to human performance in real-world scenarios.

Conclusion

This chapter provided a framework that captures the relationship between the language competence of humans and AI systems. It considers the variance between human language competence levels, while focusing only on best-performing NLP systems. The chapter analysed systems' language performance in 12 selected research areas. Results suggest that in four of these areas, machines are at the level of humans or higher; in an additional five, AI is at an advanced stage of research; and in the remaining three domains, AI is at mid-level.

The chapter discussed some challenges in directly comparing human and machine NLP capabilities. First, machines are usually developed for and tested on narrow tasks rather than broad capabilities. While there are general literacy tests for people, AI systems are evaluated on specific language tasks. This makes the comparison with respect to broader capabilities challenging. Second, machines do not have the same difficulties as humans and so understanding machine competences may require different sub-areas of language competence. Despite these difficulties, the chapter provided an initial comparison of human and AI language competence.

This work has revealed that the available data for training a system to automate a task is the factor with the most influence over progress within a specific NLP research area. Policy makers or other actors could also influence the development of AI language capabilities if they identify a new language task in which AI could have high positive impact (e.g. for governments, society or commerce). Consulting experts about what data would be needed to train systems and creating a large public dataset would allow the research community to develop systems.

In sum, as systems improve in performance, the impact of NLP technologies on society is likely to be significant. The pace of development can be very fast if large datasets become available and if there is sufficient market potential. This chapter is a first step in building an index that will help the wider public understand what current technology can do and how performance of AI in language corresponds to human language competence levels. Taking this work forward will involve the analysis of a wider set of benchmarks and the development of an AI language index that can be regularly updated to inform the education policy community on AI progress in the field.

References

Akhbardeh, F. et al. (2021), *Findings of the 2021 Conference on Machine Translation (WMT21)*. [22]

Baker, C., C. Fillmore and J. Lowe (1998), "The Berkeley FrameNet Project", *Proceedings of the 36th annual meeting on Association for Computational Linguistics -*, https://doi.org/10.3115/980845.980860. [8]

Burtsev, M. et al. (2018), "The First Conversational Intelligence Challenge", in *The NIPS '17 Competition: Building Intelligent Systems, The Springer Series on Challenges in Machine Learning*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-94042-7_2. [15]

de Marneffe, M. et al. (2021), "Universal Dependencies", *Computational Linguistics*, pp. 1-54, https://doi.org/10.1162/coli_a_00402. [9]

Dinan, E. et al. (2019), "The Second Conversational Intelligence Challenge (ConvAI2)", in *The NeurIPS '18 Competition, The Springer Series on Challenges in Machine Learning*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-29135-8_7. [16]

Dobrovolskii, V. (2021), "Word-Level Coreference Resolution", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, https://doi.org/10.18653/v1/2021.emnlp-main.605. [3]

Eric, M. et al. (2017), "Key-Value Retrieval Networks for Task-Oriented Dialogue", *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, https://doi.org/10.18653/v1/w17-5506. [19]

Göeau, H. et al. (2021), "Overview of plantclef 2021: Cross-domain plant identification". [12]

Ji, T. et al. (2022), "Achieving Reliable Human Assessment of Open-Domain Dialogue Systems", *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, https://doi.org/10.18653/v1/2022.acl-long.445. [17]

Jurafsky, D. and J. Martin (2023), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Education Inc. [10]

Koehn, P., F. Och and D. Marcu (2003), "Statistical phrase-based translation", *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, https://doi.org/10.3115/1073445.1073462. [21]

Kustritz, A. (2015), "The Fan Fiction Studies Reader ed. by Karen Hellekson and Kristina Busse", *Cinema Journal*, Vol. 54/3, pp. 165-169, https://doi.org/10.1353/cj.2015.0019. [13]

Manning, C., P. Raghavan and H. Schütze (2008), *Introduction to Information Retrieval*, Cambridge University Press, https://doi.org/10.1017/cbo9780511809071. [14]

Mrini, K. et al. (2020), "Rethinking Self-Attention: Towards Interpretability in Neural Parsing", *Findings of the Association for Computational Linguistics: EMNLP 2020*, https://doi.org/10.18653/v1/2020.findings-emnlp.65. [11]

OpenAI (2023), "GPT-4 Technical Report", *ArXiv [Cs.CL]*, http://arxiv.org/abs/2303.08774. [24]

Palmer, M., D. Gildea and P. Kingsbury (2005), “The Proposition Bank: An Annotated Corpus of Semantic Roles”, *Computational Linguistics*, Vol. 31/1, pp. 71-106, https://doi.org/10.1162/0891201053630264. [7]

Poesio, M. et al. (2016), *Anaphora Resolution*, Springer. [4]

Rajpurkar, P. et al. (2016), “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, https://doi.org/10.18653/v1/d16-1264. [18]

Srivastava, A. et al. (2022), “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models”, *arXiv preprint arXiv:2206.04615*. [23]

Sukthanker, R. et al. (2020), “Anaphora and coreference resolution: A review”, *Information Fusion*, Vol. 59, pp. 139-162, https://doi.org/10.1016/j.inffus.2020.01.010. [1]

Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, https://doi.org/10.18653/v1/w18-5446. [5]

Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. [6]

Weischedel, R. et al. (2013), “Ontonotes release 5.0”, *Linguistic Data Consortium, Philadelphia, Pennsylvania*. [2]

Xie, T. et al. (2022), “UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models”, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, https://doi.org/10.18653/v1/2022.emnlp-main.39. [20]

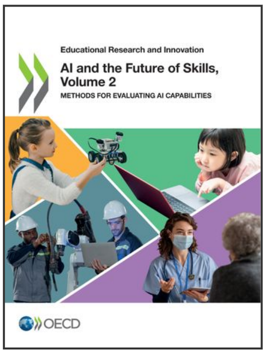
Annex 8.A. Natural Language Processing research areas

Annex Table 8.A.1. Natural Language Processing research areas with at least one benchmark task

	NLP research areas	Equivalent Human Language Competence
1	Anaphora resolution	Identify the anaphor for a specific antecedent e.g. the car (antecedent) is damaged but it (anaphor) still works.
2	Authorship verification	Identify the likelihood a text was written by a specific author or if a set of texts is likely written by the same author.
3	Automated essay scoring	Determine the academic quality of an essay.
4	Constituency parsing	Construct a constituency-based (or syntactic structure) parse tree for a sentence by applying phrase structure grammar rules.
5	Dialogue (open domain)	Carry out a conversation with another person about any topic.
6	Dialogue (task-oriented)	Assist someone in completing a task where the help provided is through conversation.
7	Emotion-cause pair extraction	Identify when a text refers to a person’s emotional state and pair that with another location a text or conversation that describes the likely cause of that emotional state.
8	Event extraction / detection	Identify the events described within a text or in a social media post.
9	Explanation generation	Explain the reasoning that led to an answer.
10	Humour detection	Identify when a sentence or paragraph would be considered humorous to some/most people.
11	Image captioning	Describe what is shown in a photo.
12	Information Retrieval	Find (and rank) the content most relevant to a specific user information need.
13	Keyword extraction	Read a text and compose a set of keywords that are likely to be used as query terms when searching for that text in a large collection of documents, such as the web.
14	Lexical normalisation	Rewrite text written in non-standard form to standard form, e.g. Fomo? fear of missing out; lol? laugh out loud; jst? just.
15	Machine Translation	Interpret, translate text or spoken language.
16	Natural Language Inference / Sentence pair modelling (entailment, semantic, similarity, paraphrase detection)	Understand the relationship between the meaning of two sentences and how the meaning of one sentence can relate in some way to the meaning of another sentence.
17	Punctuation restoration	Add the most appropriate punctuation to a text for which punctuation is not present.
18	Question Answering	General knowledge or knowledge about a specific topic; Ability to find relevant information with the help of technology, e.g. search engine.
19	Reading comprehension	Reading comprehension including the ability to identify when the text provided does not include the information required to answer a question.
20	Relation extraction	Extract the relations between entities expressed within text or conversation, e.g. Michael Jackson died in Los Angeles, CA. where diedInCity is the relation.
21	Reasoning	Reasoning.
22	Semantic parsing/role labelling	Identify the semantic roles (e.g. agent, patient) for the arguments of the predicates (e.g verb) of a sentence.
23	Sentence Compression	Rewrite a sentence as a shorter one by removing redundant information while preserving the meaning of the original sentence.
24	Sentiment Analysis (aspect-based)	Interpret the opinion being expressed within a text/conversation/social media post.
25	Sentiment Analysis (multimodal)	Interpret the opinion being expressed within multiple sources including language, e.g. facial expression and speech.
26	Speech recognition	Interpret spoken language.
27	Summarisation	Provide a written summary of a text or conversation that highlights the most important points made within a specified limit of words or sentences.
28	Text classification	Assign the text to the most appropriate category or domain (e.g. news, medical, scientific, etc.).
29	Text diacritisation	Restore the diacritics for text in languages that use diacritisation (Czech, French, Irish, etc.).
30	Text style transfer	Rewrite a text with the expressed content expressed in a specific style distinct from that of the original.
31	Topic modelling	Be able to identify the topic of a text/conversation/social media post.
32	Video captioning	Describe what is taking place in a video.
33	Visual QA	Answering a question about the content of a photo or image.

Notes

[1] Scores are an average of approximately 800 human ratings on a 100-point rating scale for a range of conversation quality criteria.



From:

AI and the Future of Skills, Volume 2

Methods for Evaluating AI Capabilities

Access the complete publication at:

https://doi.org/10.1787/a9fe53cb-en

Please cite this chapter as:

Graham, Yvette (2023), “Towards a synthesis of language capability in humans and AI”, in OECD, *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/436a79da-en

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.