Natural Language Engineering (2023), **29**, pp. 162–176
doi:10.1017/S1351324923000013

CAMBRIDGE UNIVERSITY PRESS

INDUSTRY WATCH

NLP startup funding in 2022

Robert Dale

Language Technology Group, Sydney, Australia
E-mail: rdale@language-technology.com

Abstract

It's no secret that the commercial application of NLP technologies has exploded in recent years. From chatbots and virtual assistants to machine translation and sentiment analysis, NLP technologies are now being used in a wide variety of applications across a range of industries. With the increasing demand for technologies that can process human language, investors have been eager to get a piece of the action. In this article, we look at NLP startup funding over the past year, identifying the applications and domains that have received investment.

1. Introduction

In the course of curating the content for *This Week in NLP*, a newsletter on the commercial application of NLP tools and techniques, I track company funding and acquisitions in the natural language processing space. In 2022, I found just over 340 relevant funding events, ranging from pre-seed funding all the way through to late-stage Series E and F rounds. In this article, I focus in on early-stage companies: specifically, those who reported pre-seed funding, seed funding or Series A funding rounds. Companies at this stage have yet to establish their products or services in the marketplace and can unambiguously be described as startups; they constitute the highest risk for investors, and at the same time, we might expect them to be an excellent source of innovative ideas. In the data I have available, just over 50% of the funding events were at the pre-seed, seed or Series A stage; in this article, I attempt to impose some organisation and structure over the offerings of these 173 companies, with the aim of highlighting the technology and application areas that have been considered worthy of investment over the last twelve months.

2. What's here

A flat list of 170+ companies and what they do would be quite unreadable and devoid of insight. So, the present article is organised around a taxonomisation of the space of NLP startups that is intended to be useful to those who are interested in this area. There are many ways such a space could be structured, so I'm not claiming that the set of categories used here is the only way to do this; it's simply a map of the space that makes sense to me and may make it easier for you to determine what sections of the article are most relevant to you.

In an ideal world, we might choose to organise technology offerings either by application type or, alternatively, by domain of application, viewing these as two orthogonal exhaustive solutions. But just as in my review of NLP startup funding in 2021, I think it's more informative and useful to adopt a slightly uncomfortable hybrid that picks out a hierarchy of technology types, but leaves a proportion of the companies we review as better organised in terms of domain of application. So, this article is structured as follows:

- In Section 3, we look at products and services which are primarily concerned with the processing of text, as opposed to voice: this covers the subcategories of search, information extraction, content moderation, text generation and machine translation.
- In Section 4, we look at applications that are concerned in one way or another with conversation: the subcategories identified here are development platforms, bespoke development providers, conversational intelligence, communication skills feedback, sales support and meeting productivity tools.
- In Section 5, we cover audio-visual processing, with as subcategories speech processing, voice synthesis and video synthesis.
- In Section 6, we look at domain-specific solutions, covering legal tech, health tech, ed tech and a number of loose-end domains with one or two occupants.
- Finally, in Section 7 we draw some summarising conclusions regarding directions in the field.

It would be nice if the categories outlined above were all mutually exclusive, but they're not; there's leakage and boundary crossing in various places, so you might well not agree with where I've chosen to place particular companies. In particular, there are a few companies discussed in Sections 3, 4 and 5 which could easily have been considered domain-specific applications, but in the interest of identifying clusters of activity it seemed to me more useful to include them under technology types.

Some further comments on methodology, and some caveats, are in order:

- The information presented here comes from trawling around 150 relevant news sources using a combination of manual and automatic processes. It's unlikely I captured every relevant funding event, but I believe the results are reasonably comprehensive; if you think I missed an NLP startup that received seed or Series A funding in 2022, please drop me an email so I can investigate why I missed it.
- I consider a company to be offering an NLP product or service if language processing technology appears to play an important role in that product or service. There are, inevitably, borderline cases; for example, many web-based products now incorporate a simple chatbot functionality, but I consider these in-scope only if the chatbot is an important feature or does something interesting and new.
- The descriptions I provide of what companies do are based on a brief review of each company's website at the time of writing. But things can change quickly, and companies can pivot quite significantly, so depending on when you are reading this, any given company's website might now be telling a different story. Also, I've tried to make each description as informative as possible in a short space, usually no more than one sentence; but there are a few cases where I've spent what felt like just too much time and effort on a website trying to get a clear idea of what a company's offering is, with limited success, resulting in unhappily vague descriptions.
- For each company mentioned, I indicate the year in which the company was founded, along with what series, when and how much funding was received. All amounts are in US dollars, although it should be noted that a fair number of companies are based outside the US and received their funding in other currencies; USD equivalents shown here are based on the exchange rates at the time of writing and so may vary a little from figures that were reported at the time of funding.

Anyway, enough rambling. Let's get to it.

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

3. Document processing

As the term is used in the industry, 'document AI' usually refers to approaches that have to contend with physical formatting issues in documents (like, for example, the extraction of information from tables), while 'text processing' is typically more concerned with a document's linguistic content, abstracted away from its physical rendering. My impression, however, is that an increasing number of solutions are combining these two paradigms, so I consider them here as a single combined category.

3.1 Search

A fair number of startups provide search engines that are aimed at developers who want to add search functionality to their projects. ZincSearch (founded 2022; Seed round, $3.6m, March 2022) and Meilisearch (founded 2018; Series A, $15m, October 2022) provide downloadable search engines, with Meilisearch also offering a fully managed cloud-based version. SeMI Technologies (founded 2019; Series A, $16.5m, February 2022), Hebbia (founded 2020; Series A, $30m, June 2022) and Vectara (founded 2020; Seed round, $20m, October 2022) emphasise their use of vector search, also referred to as neural search or semantic search, in contrast to older approaches based on term indexing; Pinecone Systems (founded 2019; Series A, $28m, March 2022) provides a vector database product that can serve as search infrastructure, and Nuclia (founded 2019; Seed round, $5.4m, April 2022) is an end-to-end API that lets teams use their own vectorisation and normalisation algorithms while providing storage, indexing and querying. Deepset (founded 2018; Series A, $14m, April 2022) provides an open-source NLP framework called Haystack that enables developers to build pipelines for a variety of search use cases, while Opster (founded 2019; Series A, $5m, July 2022) provides a platform for automation and management of enterprise search engines and databases.

A braver strategy is to position your search technology as an alternative to incumbents Google and Microsoft: You.com (founded 2020; Series A, $25m, July 2022), which aims to be an open search platform that allows others to build on top of its search technology, includes AI-enabled features such as YouCode, which can generate code along the lines of GitHub's Copilot based on a search query, and YouWrite, powered by OpenAI's GPT-3, which can be prompted to write essays, blog posts and boilerplate letters. But it's more common to target specific use cases: Ocean.io (founded 2017; Venture Round, $6.3m, January 2022) and Grata (founded 2016; Series A, $25m, February 2022) both aim to help enterprises find the right business targets; Vetted (founded 2019; Series A, $14m, August 2022) is a product search engine that aims to help consumers discover the brands and products most recommended for their needs; Outmind (founded 2019; Seed round, $2.1m, September 2022) focusses on aggregated search across relevant data dispersed across a range of workplace applications; Mem (founded 2021; Series A, $23.5m, November 2022) is a productivity app that searches across a user's notes; and Hypertype (founded 2021; Pre-seed round, $1.3m, May 2022) searches email archives to automate the authoring of new emails.

Outside the realm of text search, Twelve Labs (founded 2021; Seed round, $5m, March 2022) provides a video search and understanding platform which uses semantic search to locate relevant scenes across large video archives.

3.2 Information extraction

Extracting and aggregating information is a key focus for a number of startups: KnowledgeNet.ai (founded 2021; Series A, $9.4m, February 2022) aims to support deal makers and executives by integrating disjointed conversations and data across email, CRMs, file storage, professional networks and industry newsfeeds, and Ask-AI (founded 2021; Seed round, $9m, October 2022)

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

aggregates text-heavy company knowledge sources and customer communications, making the data accessible via a question-answering interface.

It's now fairly standard for general workflow automation products to include some degree of document AI capability. NanoNets (founded 2017; Series A, $10m, February 2022) lets developers create ML models which can extract data from documents and populate databases automatically; Krista (founded 2016; Series A, $15m, February 2022) emphasises the conversational nature of its low-code automation platform; and Alkymi (founded 2017; Series A, $21m, October 2022) provides a unified platform for extraction from a variety of different unstructured data sources, along with a large set of 'blueprints' for common document types.

Some information extraction products are more narrowly focussed: Neuron7.ai (founded 2020; Series A, $10m, June 2022) pitches itself as a service intelligence platform, with tech that extracts information from the data and people spread across an organisation, and uses this 'collective intelligence' to help people diagnose and resolve customer issues; Sensible (founded 2020; Seed round, $6.5m, November 2022) offers a document orchestration platform that provides pre-built templates for extracting data from over 150 insurance document types; Stimulus (founded 2017; Seed round, $2.5m, August 2022) is a relationship intelligence platform that uses data and analytics to help companies make better purchasing decisions via a proprietary scoring mechanism; Prophia's (founded 2018; Series A, $10.2m, December 2022) platform scours commercial real estate contracts and extracts key terms such as square footage and dates in the lease; and the first product from theGist (founded 2022; Pre-seed round, $7m, November 2022), theGist for Slack, provides structured, personalised summaries of Slack discussions, filtering out noise so employees don't miss important information.

3.3 Sentiment analysis

Sentiment analysis is an area that still attracts new startups, generally in the context of assessing and measuring consumer or user feedback. A common focus is the aggregation and categorisation of feedback across multiple channels or sources, these days using AI models: Viable (founded 2020; Seed round, $5m, May 2022) has GPT-3 under the hood and offers a full service that provides written analysis on top of the aggregated feedback, as well as natural language querying of that feedback; Idiomatic (founded 2016; Seed round, $4m, May 2022) categorises feedback using models that are tailored to each specific business case; and Spiral (founded 2018; Seed round, $1.3m, November 2022) sells its feedback tech to medium-to-large companies in the banking, fintech, connected devices and insurance industries. Lang (founded 2018; Series A, $10.5m, May 2022), Unwrap (founded 2021; Seed round, $3.2m, July 2022) and Sturdy AI (founded 2019; Seed round, $3.1m, June 2022) all similarly categorise detected issues and concerns and provide various forms of analysis.

A related focus is brand management: My Telescope (founded 2018; Pre-seed round, $2.6m, March 2022) is a market intelligence and search platform for marketers and brands that predicts the long-term impact of market trends, brand strength and campaign effectiveness, and Knit (founded 2015; Seed round, $3.6m, June 2022) uses a network of young consumers to provide brands with detailed consumer insights based on video feedback and quantitative surveys; the company's video analysis AI claims to analyse hours of video feedback in minutes.

3.4 Content moderation

Although the technology categories discussed above are somewhat traditional and long-standing, 'content moderation' is a category that has emerged only in recent years and is set to grow as increasing attention is paid to issues like disinformation and harmful language use.

Fairwords (founded 2014; Series A, $5.3m, February 2022) uses a spell-check like interface to notify users of harmful language as they type and give information about how the language

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

might be interpreted; the software can also look for signs of bribery and corruption, collusion and discrimination; mpathic (founded 2021; Seed round, $4m, June 2022) similarly helps employees identify potential misunderstandings or misinterpretations in their communication and adjusts in real-time; Checkstep (founded 2020; Seed round, $5m, May 2022) targets misinformation, hate speech, Child Sexual Abuse Material (CSAM), bullying and spam, and also has copyright infringement management capabilities; Areto Labs (founded 2020; Pre-seed round, $730k, June 2022) helps companies identify and monitor online abuse, and addresses instances of it through automated counteractions like muting, blocking and reporting the accounts responsible; and Modulate (founded 2017; Series A, $30m, August 2022) is the developer of ToxMod, a tool for detecting and dealing with violent or otherwise offensive speech in real-time video game voice chat. Diversio (founded 2018; Series A, $6m, January 2022) measures and tracks language around diversity, equity and inclusion, identifying 'inclusion pain points' in employee-authored text.

VineSight (founded 2018; Seed round, $4m, September 2022) tracks, analyses and mitigates online misinformation and toxicity against brands, campaigns and causes; Alethea (founded 2019; Series A, $10m, November 2022) detects and mitigates instances of disinformation and social media manipulation; Logically (founded 2017; Series A, $24m, March 2022) combines AI with expert analysts to find, triage and respond to information threats; and Pendulum's (founded 2021; Seed round, $5.9m, January 2022) platform uses 'narrative tracking' across a range of media to uncover threats and opportunities contained in narratives in the earliest days of their formation, tracking them as they spread online.

A related area is privacy management: Redactable (founded 2018; Pre-seed round, $1.2m, May 2022) and Private AI (founded 2019; Series A, $8m, November 2022) automatically detect Personally Identifiable Information (PII) in documents and redact it; Lightbeam.ai (founded 2020; Seed, $4.5m, April 2022) attempts to identify which particular customer or identity the information belongs to so security teams can automate the protection of that data more effectively; and Protopia AI (founded 2020; Pre-seed round, $2m, December 2022) focuses on risks while data are in use during ML inference, obfuscating personal information so that it isn't identifiable or at risk of leakage to unauthorised third parties.

Another related area is risk management. Shield (founded 2018; Series A, $15m, January 2022) produces a workplace intelligence platform for compliance teams: its tech applies NLP to employee communication channels to detect conduct violations such as market manipulations; Concentric AI (founded 2018; Series A, $14.5m, May 2022) identifies and categorises sensitive information, assesses risk and resolves security issues via a measure called 'Risk Distance'; and VISO Trust (founded 2020; Series A, $11m, March 2022) is a security due diligence platform which automates the process of compiling third-party cyber risk data using document heuristics, NLP and ML. These are company-internally focussed solutions; KYP (founded 2021; Pre-seed round, $960k, October 2022), on the other hand, is a third-party risk intelligence platform that aims to deliver a complete picture of the partners a business relies on.

3.5 Text generation

You'd have to be living under a rock to have missed the recent intensity of interest in generative AI and the use of large language models for text prediction in particular. This year's most visible startup in this space, thanks to the size of its most recent funding round, is content platform Jasper (founded 2021; Series A, $125m, October 2022); this was the biggest single NLP startup funding event I'm aware of in 2022. There's also Regie.ai (founded 2020; Seed round, $4.8m, June 2022), whose GPT-3-powered copy-writing platform focuses on sales and marketing teams.

Solutions that appear to be based on older approaches to text generation are still attracting funding too: Linguix (founded 2018; Pre-seed round, $1m, February 2022) is a writing assistant that offers spelling and grammar checking as well as text rephrasing and various scoring metrics; and Magical (founded 2020; Series A, $35m, June 2022) is a productivity aid akin to

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

a text expander, where the software detects elements on web pages and enables the creation of custom abbreviations for moving the corresponding text around. QorusDocs (founded 2012; Venture round, $10m, October 2022) is cloud-based proposal management software that streamlines RFP responses and automates proposal creation; the software leverages NLP to streamline the RFP response process by selecting the most important and relevant content from a company's document archives.

Also related here are Mintlify (founded 2020; Seed round, $2.8m, May 2022), whose platform reads code and creates documentation to explain it and also detects how users engage with the documentation to improve its readability, and Findable (founded 2020; Seed round, $2.1m, June 2022), whose tech automates the organisation of documentation for buildings by analysing headlines, pictures and drawings.

3.6 Machine translation

Language I/O (founded 2011; Series A, $6.5m, January 2022) provides a translation platform that allows customers to provide real-time customer support in more than 100 languages; Viva Translate (founded 2020; Seed round, $4m, February 2022) is a cross-language translation tool that focuses on translation in the context of freelancer-client communications; Weglot (founded 2016; Series A, $48m, March 2022) is a no-code website localisation tech provider whose platform supports human refinement via post-editing features; XL8 (founded 2019; Pre-Series A, $3m, July 2022) provides MT technology optimised for media content, including synthetic dubbing or voice-over; and WritePath (founded 2009; Seed, $340K, December 2022) is a cloud-based B2B translation platform targeting business, ESG and investor relations disclosure.

3.7 Miscellaneous other applications

There are a number of other companies that process natural language textual input in one way or another, but which don't fit comfortably into the already-stretched categories above. In the text-to-image space, there's visual art startup Stability AI (founded 2019; Seed round, $107m, October 2022), the company behind Stable Diffusion. Spiritt (founded 2020; Pre-seed round, $5.5m, July 2022) transforms a textual description into an app, gathering the information required via a conversation with a chatbot. Zenlytic (founded 2018; Seed round, $5.4m, November 2022) is a no-code business intelligence tool that provides a natural language interface. And Unlikely AI (founded 2018; Seed round, $20m, September 2022), who tease with the claim to be pursuing an alternative to large neural networks, launched as their first product an app that solves and explains cryptic crossword clues.

4. Conversational AI

4.1 Development platforms

It appears there's still space in the market for new self-service conversational AI development platforms. Some of these focus on text-based chatbot development: Druid (founded 2018; Series A, $15m, May 2022) and OpenDialog AI (founded 2019; Seed round, $4.8m, May 2022) offer no-code chatbot authoring platforms; Zowie (founded 2019; Seed round, $5m, January 2022) targets businesses who sell online, combining its no-code automation capabilities with a suite of tools that allows agents to provide personalised care and product recommendations.

Others add voice capabilities: NLX (founded 2018; Seed round, $5m, January 2022) and Parloa (founded 2017; Seed round, $4.25m, May 2022) offer no-code/low-code platforms for automating omnichannel customer service that includes phone and chat, and Flip CX (originally RedRoute; founded 2017; Seed round, $6.5m, February 2022) emphasises the importance of being able to handle voice calls, providing an easy-to-use configuration tool that leverages already-designed call flow patterns.

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

4.2 Bespoke development

There are also a good number of new players who will build you a conversational app using their own platforms and toolsets. Futr (founded 2017; Seed round, $2.5m, April 2022) emphasises its platform's ability to support multilingual live chat on all social channels; Tenyx (founded 2021; Seed round, $15m, May 2022) builds voice-based virtual customer service agents using what it calls 'neuroscience-inspired' AI; Curious Thing (founded 2018; Seed round, $4.7m, May 2022), whose tech previously focused on interactions around HR, has pivoted to providing a broader range of voice-driven conversational AI solutions covering both inbound and outbound calls; and Tymely (founded 2020; Seed round, $7m, September 2022) uses AI-human hybrid tech to automate customer service capabilities, with each machine-generated response being verified by a human agent.

Chatdesk (founded 2016; Series A, $7m, January 2022) has an interesting model: eschewing chatbots altogether, it finds, hires and trains 'superfans' of a brand to become 'Chatdesk Experts', and on the back-end, uses ML to analyse previous support messages, creating an on-brand knowledge base that enables these superfans to respond to customer questions with the brand's voice and policies.

4.3 Conversational intelligence

Continuing a trend that's been visible for a few years now, a number of companies offer tech that carries out some form of analysis of conversational interactions, whether those involve virtual or human agents.

Wiz.ai (founded 2019; Series A, $20m, January 2022), focussing on conversational AI for Southeast Asian languages, uses its front-end talkbot to encourage customers to engage in conversations, while the back-end sifts through data in real time and stores insights from the conversations into the company's existing CRM system for subsequent analysis; Talkmap (founded 2017; Series A, $8m, February 2022) labels, structures and analyses interactions with customers with the aim of providing near real-time insights into those conversations; Affogata (founded 2018; Seed round, $9.5m, March 2022) provides a voice analysis platform that allows businesses to identify unusual patterns, to streamline real-time responses and engage in pre-emptive action; Winn.AI (founded 2021; Seed round, $17m, September 2022) monitors sales calls to automatically track, capture and update CRM entries, reducing the need for salespeople to take notes themselves; and Operative Intelligence (founded 2021; Seed round, $3.5m, December 2022) provides tech that is designed to help call centre operators overcome mismatched perceptions about why customers contact them, reducing wait times and improving issue resolution by identifying the real reasons. Jiminny (founded 2016; Series A, $17m, August 2022) is a conversational intelligence platform that analyses emotion in video, automates the scoring of call interactions and generates real-time insights.

4.4 Communication skills feedback

The analysis of a human agent's conversational contributions in order to provide feedback on communication skills can be thought of as a specific form of conversational intelligence.

Abstrakt (founded 2020; Pre-seed round, $120k, March 2022) provides real-time call coaching that listens in on a call and makes helpful suggestions; Klaus (founded 2017; Series A, $12m, September 2022) coaches agents by keeping track of a range of communicative KPIs, identifying coaching opportunities and measuring support quality.

Call Simulator (founded 2021; Seed round, $575k, January 2022) is a conversational simulation platform that aims to prepare call centre agents for real-world scenarios; Second Nature (founded 2018; Series A, $12.5m, January 2022) hosts a simulator with avatars that have conversations with sales reps, measuring how deeply the reps cover key topics.

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

More broadly, Yoodli (founded 2021; Seed round, $6m, August 2022) analyses speech to offer tips for improving communication skills: the platform provides users with a transcript and analysis on use of filler words, non-inclusive language, pacing, body language and other actionable insights. The company recently struck a deal to provide speech coaching for Toastmasters International, the well-known public speaking and leadership training organisation.

4.5 Sales support

There are also a number of companies offering various forms of what we'll refer to here as sales support. Tactic (founded 2020; Seed round, $4.5m, March 2022) automates customer and market research by allowing sales and marketing staff to ask questions of customer and market data in ordinary language and to apply filters for prioritising and ranking the results; Connectly.ai (founded 2020; Seed round, undisclosed amount, July 2022) is a no-code tool that allows businesses to create and send interactive and personalised marketing campaigns through AI-powered 'mini bots'; Demoleap (founded 2020; Seed round, $4.4m, August 2022) is a live demo assistant and sales discovery platform that guides sellers in following the sales process throughout a live demo; Heyday (founded 2021; Seed round, $6.5m, June 2022) is a conversational AI platform for retailers that automates FAQs; and AdTonos (founded 2016; Seed round, $2.1m, August 2022) monetises audio streams via interactive ads played through smart speakers and mobile devices using its YoursTruly platform.

4.6 Meeting productivity tools

The COVID-driven increase in use of virtual meeting platforms like Zoom and Teams has created a market for a relatively new category of tools that aim to support meeting productivity; in many regards, these are a repurposing of tools and techniques developed in the context of conversational intelligence.

Sembly AI (founded 2019; Seed, undisclosed amount, March 2022), Headroom (founded 2020; Seed round, $9m, August 2022), Xembly (founded 2020; Series A, $15m, October 2022), Fathom (founded 2020; Seed, $4.7m, November 2022) and tl;dv (founded 2020; Seed round, $4.6m, June 2022) all offer some combination of functionalities for transcribing and analysing meetings, extracting topics and action items, and generating summaries and meeting minutes.

Airgram (founded 2020; Series A, $10m, August 2022) is an audio-visual recording tool that can be set to auto-join a pre-scheduled Zoom, Google Meet or Microsoft Teams meeting, recording it when the user isn't present; the tool provides flexible playback options along with transcription and detection of topics and action items; and Amy (founded 2019; Seed round, $6m, June 2022) is a sales intelligence platform that aims to help streamline meeting preparation by leveraging publicly available information about a prospect, transforming this data into meeting briefs that provide tangible insights into the prospect.

5. Audio-visual processing

We introduce this category to cover applications where voice technology is used other than in support of conversational AI, and also where it's used in combination with video.

5.1 Speech processing

NeuralSpace (founded 2019; Seed round, $1.7m, February 2022) specialises in the development of voice technology for low-resource languages, providing a self-service toolkit that covers more than 90 languages and incorporates automatic language detection; Ava (founded 2014; Series A, $10m,

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

March 2022) is a live captioning platform that listens to audio during meetings or from videos to provide captioning for the deaf and hard of hearing, tagging each caption with its speaker; Sounder (founded 2019; Series A, $7.7m, February 2022) is an end-to-end podcast management platform that incorporates brand safety and brand suitability analysis, topic analysis, content summarisation, and dynamic segmentation; and AssemblyAI (founded 2017; Series A, $28m, March 2022) provides a set of LLM-based 'audio intelligence' APIs for transcribing and understanding audio data, with applications including content moderation, emotion detection, summarisation, and PII redaction.

Sanas (founded 2020; Series A, $32m, June 2022) provides real-time accent translation, assisting multilingual speakers to deliver clear communication through accent correction, and Namecoach (founded 2014; Series A, $8m, November 2022) provides software that embeds context-aware audio name pronunciation buttons in the applications people use every day, enabling users to pronounce names with confidence.

5.2 Voice synthesis

Murf AI (founded 2020; Series A, $10m, September 2022) is a synthetic speech technology startup developing lifelike AI voices for podcasts, slideshows and professional presentations, with a curated voice library of over 120 voices in more than 20 languages. At the level of specific applications, ping (founded 2016; Seed round, $5m, June 2022) allows commercial drivers to hear their smartphone messages and emails read out loud in more than 105 languages.

A major use for voice synthesis is in audio dubbing into other languages. Dubverse (founded 2021; Seed round, $800k, June 2022) is an automated dubbing platform that allows users to dub videos into multiple languages in almost real time, currently available in 10 Indian languages and 20 'world' languages; Dubdub (founded 2021; Seed round, $1m, September 2022) uses AI and ML to create multilingual video content for businesses across 40 languages; Deepdub (founded 2019; Series A, $20m, February 2022) provides a dubbing service for entertainment content, using synthesised versions of the original actors' voices so the dubbed version sounds more like the original; and Papercup (founded 2017; Series A, $20m, June 2022) similarly translates videos by generating voices that sound like the original speaker. These applications typically provide a human-in-the-loop feature, whereby a professional translator can perform a quality check, editing and amending translation and speech to improve quality.

5.3 Video synthesis

We include here companies whose tech focusses on the creation of video output, since these typically also involve voice synthesis.

Pictory (founded 2019; Seed round, $2.1m, January 2022) converts long-form content such as webinars, blogs and white papers into short social videos, and ShortTok (founded 2021; Pre-seed round, undisclosed amount, October 2022) develops automated visual storytelling technologies that create short form videos from a client's library of video and multimodal content. Peech (founded 2020; Seed round, $8.3m, August 2022) offers a video editing tool for content marketing teams which automatically synthesises on-brand visuals to match content and cuts filler words; Rephrase.ai (founded 2019; Series A, $10.6m, September 2022) also builds generative AI tools for synthetic video creation for marketing and content teams.

A particular form of video synthesis is the generation of virtual characters. Metaphysic (founded 2021; Seed round, $7.5m, January 2022), the company known for its Tom Cruise deepfake, develops tools for creating digital avatars that can integrate into the metaverse; Inworld AI (founded 2021; Seed round, $12.5m, March 2022) is another platform for creating AI-driven virtual characters, immersive realities and metaverse spaces, enabling non-technical users to create character personalities by describing them using natural language; and Deep Voodoo

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

(founded 2020; Seed round, $20m, December 2022) is a deepfake startup launched by South Park creators Trey Parker and Matt Stone.

Speech Graphics (founded 2010; Series A, $7m, February 2022) offers audio-driven facial animation technology that enables animated characters in games and other applications to move their mouths correctly when speaking; Hour One's (founded 2019; Series A, $20m, April 2022) tech converts people into virtual human characters that can be activated with lifelike expressiveness. Carter (founded 2022; Seed round, $2m, December 2022) is working on conversational AI to help games developers make computerised gaming characters more lifelike. NeuralGarage (founded 2021; Seed round, $1.5m, November 2022) is a video dubbing platform: given an audio input and a human face, it transforms the lip and jaw movements of the person to match the speech, irrespective of language.

6. Domain-specific solutions

6.1 Legal tech

The law is intimately bound up with language, so it's not surprising that legal tech has long been an important domain for the application of language processing techniques.

A popular area is document analysis and review, where AI-supported analysis can reduce the significant amount of time typically spent on manual processing. TermScout (founded 2018; Seed round, $5m, May 2022) extracts key information from contracts to enable easier review, rating and comparison against industry standards; Terzo (founded 2020; Series A, $16.3m, November 2022) extracts critical data from contracts to help organisations optimise their spend and revenue across their supplier and customer relationships; Nammu21 (founded 2017; Series A, $15.8m, October 2022) deconstructs loan documents into structured data; Summize (founded 2018; Series A, $6m, October 2022) is a contract review solution that aims to improve collaboration between in-house legal and business users by integrating with Teams, MS Word and Slack; and Della (founded 2018; Seed round, $2.5m, March 2022) focuses on complex single documents in contrast to large document review projects. Zero (founded 2014; Series A, $12m, March 2022) is an iOS mobile device-based productivity aid that integrates with email inboxes and document management systems, extracting key information such as billable interactions and automatically filing emails into folders.

Another popular area is the provision of support in drafting legal documents. Henchman (founded 2020; Seed round, $3.2m, February 2022) is a contract drafting startup that provides a Microsoft Word plug-in that suggests clauses from a firm's database as you work; LexCheck (founded 2015; Seed round, $5m, March 2022) provides a contract negotiation solution that analyses contracts to build issues lists and revisions to contract language; and Harvey (founded 2022; Seed round, $5m, November 2022) uses GPT-3 to draft documents for lawyers given a description of the task to be accomplished; the application can also answer legal questions.

A number of companies combine both these review and drafting functionalities together with other activities to provide more comprehensive legal automation platforms. Uhura Solutions (founded 2018; Seed round, $1.8m, April 2022) is a low-code contract intelligence platform that uses NLP to streamline the process of analysing and drafting contracts and agreements; Goodlegal (founded 2021; Pre-seed round, $1.3m, November 2022) offers a set of automation tools that includes a drag-and-drop editor for building legal texts, with the ability to check every legal text produced against legally compliant standards; PocketLaw (founded 2018; Series A, $10.6m, May 2022) is a contract automation SaaS legal tech platform mainly focused on SMEs; Klarity (founded 2017; Series A, $18m, January 2022) provides an automated document processing and management platform for finance and accounting teams; and Josef (founded 2017; Seed round, $5.2m, November 2022) is a no-code software platform that allows legal professionals to automate repetitive tasks including document drafting, offering legal guidance and advice and building bots for

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

client interviews. Legal OS (founded 2018; Seed round, $7m, January 2022) is a no-code legal automation platform that turns expert knowledge into a digital knowledge graph which can later be used to build a variety of legal products and processes.

There are a few other legal tech solutions that don't fit neatly into the above categories. Alchemy Machines (founded 2021; Pre-seed round, $400k, March 2022) uses NLP and speech recognition to transcribe, analyse and summarise legal-specific web meetings and phone calls; Neur.on (founded 2022; Seed round, $1.7m, August 2022) provides custom MT solutions for legal professionals; Ex Parte (founded 2017; Series A, $7.5m, February 2022) uses ML to predict the outcome of litigation, recommending actions that its customers can take to optimise their odds of winning; and Proof Technology (founded 2017; Series A, $5.5m, March 2022) is a rather unique end-to-end solution that analyses court documents to extract case caption information, determines the location of the closest process server to the defendant's or witness's address, prints the relevant material remotely and captures photographic and descriptive data about the attempt to serve.

6.2 Health tech

Another domain that has had a long relationship with NLP is that of health tech. Two key areas here are the use of conversational AI in healthcare settings and the processing of medical records.

On the conversational AI front, HeyRenee (founded 2021; Seed round, $4.4m, January 2022) is a patient-centric personal health concierge that can remind users about pills they need to take, monitor health vitals, handle prescription refills and set up virtual or in-person visits with doctors; Apowiser (founded 2021; Seed round, $1.5m, June 2022) makes PharmAssist, a chatbot-based system to support customers in the online purchase of Over-the-Counter drugs, identifying issues that should be escalated to a healthcare provider and checking for allergies and potential sensitivities to ingredients of OTC medications; BirchAI (founded 2020; Seed round, $3.1m, January 2022) aims to streamline customer support for healthcare companies by summarising and analysing the contents of phone conversations between customers and representatives; WhizAI (founded 2017; Series A, $8m, September 2022) provides a conversational interface to an analytics platform targeted at the life sciences and healthcare industries; and Kahun (founded 2018; Seed round, $8m, September 2022) makes a clinical assessment chatbot, built on the company's proprietary map of over 30 million evidence-based medical insights.

With regard to medical record processing, DigitalOwl (founded 2017; Series A, $20m, January 2022) makes a platform for medical records analysis which extracts relevant information from large collections of documents; Dyania Health (founded 2019; Seed round, $5.3m, September 2022) makes an NLP platform that performs disease-specialised clinical text extraction; Wisedocs (founded 2018; Seed round, $3m, March 2022) uses Intelligent Character Recognition to read and analyse various documents involved in a medical record review; and XpertDox (founded 2015; Seed round, $1.5m, August 2022) is the maker of XpertCoding, a tool that utilises AI to autonomously code medical claims. DeepScribe (founded 2017; Series A, $30m, January 2022) is an ambient medical scribe that records doctor-patient conversations, summarising and integrating them into the health record system; Abridge (founded 2018; Series A, $12.5m, August 2022) is a conversational AI startup that structures and summarises medical conversations for doctors and patients, helping to fill in health records with relevant information; and Eleos Health (founded 2019; Series A, $20m, April 2022) builds clinical applications that operate ambiently in the background of behavioural health clinician-patient conversations, generating post-session clinical progress notes and insurance coding.

Finally, a few other health tech startups that don't fit into the above categories: Kintsugi (founded 2019; Series A, $20m, February 2022) detects signs of clinical depression and anxiety using machine learning and voice biomarkers; Marigold Health (founded 2016; Seed round, $6m, February 2022) uses chat support groups to help individuals in recovery from a substance use

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

or mental health condition, with NLP aiding peers in moderating their online community; and WeWalk (founded 2019; Grant, $2m, July 2022) is a startup developing a smart cane for the visually-impaired: its voice assistant can answer questions about where the user is and nearby public transport, identify nearby buildings and landmarks, book an Uber and provide real-time walking directions in a form suited to those with limited or no vision.

6.3 Other domains

A third domain that has seen some interest is educational technology. FoondaMate (founded 2020; Seed round, $2m, May 2022) is a chatbot that makes education accessible to students in developing countries by letting them ask questions via Facebook and WhatsApp; Prof Jim (founded 2020; Seed round, $1.1m, January 2022) works with textbook publishers and education providers to convert textbooks and other text-based learning material into online courses including automatically generated assessments and avatar instructors; Language Confidence (founded 2016; Seed round, $1.5m, March 2022) makes an API that listens to students and assesses and corrects their spoken English pronunciation with visual feedback; and Copyleaks (founded 2015; Series A, $6m, May 2022) is a plagiarism detection solution that identifies and tracks plagiarised content online in 100+ languages.

Then we have a collection of other targeted applications in a variety of other domains:

- Financial services: Aviva (founded 2022; Pre-seed round, $2.2m, December 2022) uses NLP to match customers' spoken words to the fields of a real-time credit application; Webio (founded 2016; Series A, $4m, June 2022) provides credit, collections and payment businesses with a no-code conversational AI platform that allows customers to ask questions, change payment dates, or organise a new repayment schedule.
- Real estate: DOSS (founded 2015; Seed round, undisclosed amount, October 2022) provides a conversational assistant that lets customers ask for real-estate advice and tips, search for home listings and get neighbourhood information and recent sales data.
- DevOps: Kubiya (founded 2022; Seed round, $6m, October 2022) makes a conversational AI solution for DevOps teams, allowing users to express their intent in natural language and have the virtual assistant automate simple and tedious tasks.
- Deskless work environments: Datch (founded 2018; Series A, $10m, July 2022) operates as an intelligent voice interface in industrial environments.
- Retail: Evabot (founded 2016; Series A, $8.3m, July 2022) curates ideas for corporate gifts, suggesting what might best suit the user's clients through a chatbot-administered questionnaire; it also utilises GPT-3 to compose a personalised note that is 'handwritten' by a machine holding a pen for each gift.
- Restaurants: Valyant AI (founded 2017; Seed round, $4m, April 2022) develops a proprietary conversational AI platform for the restaurant, retail and service-based industries, and ConverseNow (founded 2018; Series A, $10m, August 2022) provides a conversational platform for restaurants to automate the order-taking process from high-volume voice channels.
- Children: Snorble (founded 2019; Seed round, $10m, April 2022) makes a bedtime robot for kids that incorporates a voice-driven assistant that can tell stories, take the child through breathing exercises and play calming music accompanied by a light show.

7. Summing up

So there you have it: 173 companies that received startup funding in 2022 for tools and applications that leverage NLP techniques, for a total investment value of just over US$1.8 billion. Table 1

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

Table 1. Funding breakdown by category and subcategory

Major category	Minor category	Investment	% of Major	% of Total
Document processing	Search	$239.7m	26.5%	12.9%
	Information extraction	$100.6m	11.1%	5.4%
	Sentiment analysis	$33.3m	3.7%	1.8%
	Content moderation	$152.1m	16.8%	8.2%
	Text generation	$180.7m	19.9%	9.8%
	Machine translation	$61.8m	6.8%	3.3%
	Misc other text	$137.9m	15.2%	7.4%
Category subtotal		$906.1m		48.9%
Conversational AI	Development platforms	$40.6m	11.0%	2.2%
	Bespoke development	$36.2m	9.8%	2.0%
	Conversational intelligence	$75.0m	20.3%	4.1%
	Comms skills feedback	$151.1m	40.9%	8.2%
	Sales support	$17.5m	4.7%	0.9%
	Meeting productivity	$49.3m	13.3%	2.7%
Category subtotal		$369.6m		20.0%
Voice and video	Speech processing	$87.4m	37.1%	4.7%
	Voice synthesis	$56.8m	24.1%	3.1%
	Video synthesis	$91.5m	38.8%	4.9%
Category subtotal		$235.7m		12.7%
Domain-specific	Legal tech	$129.8m	38.2%	7.0%
	Health tech	$145.3m	42.7%	7.8%
	Ed tech	$10.6m	3.1%	0.6%
	Other domains	$54.5m	16.0%	2.9%
Category subtotal		$340.2m		18.4%
Overall total		$1851.7m		

shows the investment breakdown in terms of the categories used here to structure the space, but do remember the various caveats about coverage and categorisation in Section 2. At best, we might view this breakdown as being broadly indicative of where the action is.

For me, a number of things stick out here. Here are my top 10 takeaways from this exercise.

(1) Despite Google's near monopoly of the search space – it's typically reported as having over 80% of search engine usage – investors still think search is worth pumping money into. Much of the activity here doesn't compete directly with Google, being more oriented towards enterprise search and other specific uses cases, but just as this piece was being

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

written, it was reported that Google's management had declared a 'code red' in response to the appearance of OpenAI's ChatGPT question-answering chatbot, and a worry that such approaches could reinvent or replace the traditional internet search engine. Notably, You.com was quick to integrate a ChatGPT-like model and interface into its search engine; Perplexity.ai and Neeva are also experimenting with combining traditional search with LLMs.

(2) Document AI appears to have enjoyed a recent revival thanks to the application of deep learning techniques, and an increasing number of applications signal a recognition that real solutions in the document processing space can't restrict themselves to disembodied text and have to take a 'whole of document' stance in order to generate value. I suspect we'll see further advances in the combination of Intelligent Character Recognition and LLM-based text processing, although it's still early days: I had occasion earlier in the year to experiment with a number of document AI products and was disappointed at the difficulties they had in reliably extracting information from what might be considered fairly simple tables.

(3) Sentiment analysis broadly construed is concerned with pragmatic aspects of language use. We've come a long way from the early days of determining the polarity of movie reviews or product recommendations, with variations of sentiment analysis techniques generating significant interest in two key areas that nobody recognised 10 years ago: the analysis of efficacy in communication (whether in sales calls or in meeting contexts), and content moderation. The second of these in particular seems ripe for growth, in the face of concerns about the divisiveness and nastiness of social media platforms, and the impact of regulations like Europe's Digital Services Act.

(4) Commercial applications of text generation are on the cusp of a dramatic paradigm shift, with the template-based technologies prominent in the last decade now appearing so mundane in comparison to recent advances in LLM-based text generation that they struggle to be considered a part of AI. Those earlier approaches will still have a role for the foreseeable future, but they are increasingly far from the cutting edge. Of course, we still have to overcome the problem that large language models are not reliable sources of truth.

(5) Machine translation feels like it's a mostly-solved problem or at least a problem that you'd want to think twice about before investing in. There's always room for incremental improvements, but those are likely to come from competition between the big incumbents like Google and Microsoft; from an investor's point of view, the future lies in packagings of functionality in interesting configurations with other technologies, or to address interesting and innovative use cases.

(6) Conversational AI in general is an area that confounds me: it's a very densely populated space, and one in which it's often hard to see what any given company's unique selling point is. I'd hate to be in the market for a development tool or application developer in this space, and I struggle to see what motivates investment decisions here. Perhaps the likes of ChatGPT will shake things up in this space, although again there's that gnarly caveat about the need to speak the truth. For the time being, I think the scope for innovative developments here lies in the analysis of conversations in service of goals like the identification of issues, coaching around communication skills, and other uses we haven't thought of yet.

(7) There's not much to be said about COVID that's positive, but it certainly gave a boost to makers of online meeting tech apps like Zoom, Microsoft Teams and Google Meet, and that in turn opened up a whole new area of tech in the form of meeting productivity tools. As noted above, these are often multi-party variants of tools developed for conversational intelligence, but it's an area that, given the nature of the content it deals with – what we might call 'long-form dialog' – offers opportunities for the exploitation of just about every type of NLP technology. I think this is one area to keep a close eye on.

https://doi.org/10.1017/S1351324923000013 Published online by Cambridge University Press

(8) Compared to a few years ago, voice synthesis has now reached a very high standard, and I think it's an area where we can only expect incremental improvements. Voice dubbing is now where it's at, particularly in combination with synthetic lip movements in video; the potential impact of convincingly realistic audio-visual translation across the entire catalogues of the likes of Netflix is immense.

(9) Legal tech and health tech have been key application areas for NLP for a while, and they are likely to remain so. It's interesting – and, from where I stand, a bit worrying – that LLMs are finding their way into legal tech, and health tech is sure to follow. I hate to sound like a scratchy record, but if there's a reckoning coming, this is where it will come into view: It's only a matter of time before a fatigued or inattentive human editor lets through a gross unfactuality in an LLM-drafted contract, leading to an expensive remediation, or – worse – in a health report, leading to injury or even loss of life. We might expect to see proponents argue that, analogously to the claims made for self-driving cars, the use of LLMs in these contexts nonetheless leads to fewer errors and less damage overall.

(10) How will the education sector deal with large language models? ChatGPT provoked a fair bit of media coverage around the theme of 'the end of the college essay', with the fear that ghost essay writing services would become easily accessible to even the poorest students, and just as this article was being written, there was a flurry of coverage over a South Carolina professor's discovery of a student who used the app to write a philosophy essay. This will be interesting to watch: either the tech will develop in directions that address traditional educational concerns or traditional education assessment practices will have to radically change.

Overall, 2022 saw the introduction of some fascinating technology and solutions. With GPT-4 just over the horizon, 2023 promises to be an even more interesting year.

If you'd like to keep up to date with what's happening in the commercial NLP world, please consider subscribing to the free *This Week in NLP* newsletter at https://www.language-technology.com/twin.

Cite this article: Dale R (2023). NLP startup funding in 2022. *Natural Language Engineering* **29**, 162–176. https://doi.org/10.1017/S1351324923000013