Natural Language Engineering (2019), **25**, pp. 405–418
doi:10.1017/S1351324919000081

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Anniversary article: Then and now: 25 years of progress in natural language engineering

John Tait[1,*] and Yorick Wilks[2]

[1]Johntait.net Ltd, Thorpe Thewles, Stockton-on-Tees, UK and [2]Florida Institute of Human and Machine, Cognition 15, SE Osceola, Ocala FL 34471, USA
*Corresponding author. Email: john@johntait.net

Abstract

The paper reviews the state of the art of natural language engineering (NLE) around 1995, when this journal first appeared, and makes a critical comparison with the current state of the art in 2018, as we prepare the 25th Volume. Specifically the then state of the art in parsing, information extraction, chatbots, and dialogue systems, speech processing and machine translation are briefly reviewed. The emergence in the 1980s and 1990s of machine learning (ML) and statistical methods (SM) is noted. Important trends and areas of progress in the subsequent years are identified. In particular, the move to the use of n-grams or skip grams and/or chunking with part of speech tagging and away from whole sentence parsing is noted, as is the increasing dominance of SM and ML. Some outstanding issues which merit further research are briefly pointed out, including metaphor processing and the ethical implications of NLE.

Keywords: Information extraction; Machine learning; Machine translation; Parsing; Statistical methods

1. Introduction

In this paper, we begin by reviewing the state of the art of practical natural language processing (NLP), or computational linguistics, or natural language engineering (NLE) (or whatever you prefer), around the time of the publication of the first volume of *Natural Language Engineering* (*NLE*) in 1995. The focus is on text processing, which has always been the primary focus of this journal in practice, although speech processing always got some consideration. We also concentrate on material published in *NLE*.

We go on to review the state of the art at the time of writing in late 2018, and then move to areas where progress has not been made, or not made in the way expected, over the past 25 years.

Finally, we identify some issues which, in our view, merit more and better research and engineering.

The reasons why the term 'Natural Language Engineering' was selected as the title of the journal have been discussed elsewhere (Boguraev, Garigliano and Tait 1995; Tait 2019). This problem of terminology and, therefore, the scope of the journal has been debated on the pages of the journal. For example, Cunningham (1999) contrasts the term 'Language Engineering' with 'Computational Linguistics' and 'Natural Language Processing'. A similar analysis appears in Thompson (1983), which we will return to in Section 2.1.

One of the authors (JT) has been associated with the *NLE* project since inception (see Tait 2019) as an editor and executive editor and has been active in computer processing of natural language since the 1970s. The other (YW) was a founding member of the Editorial Board of *NLE* and has been active in the field since the 1960s.

2. State of the art in the early 1990s

In the early 1990s, NLP – leaving aside speech processing and machine translation (MT) for the moment – was dominated by two main strands. One was the parsing of natural language (usually English) texts using large, hand-coded grammars together with linguistic resources, often hand-coded as well. The other was Information Extraction (IE), essentially template filling, following the model of the Message Understanding Conferences (MUC) (Grishman and Sundheim 1996), in which pre-specified templates (essentially structured database entries) were populated by analysis of (again, usually English) natural language texts. These were generally shallow analysis systems, without any attempt at full syntactic parsing. There was also a third body of work on the modelling of dialogue, using methods drawn from models of interaction and belief, but avoiding the methods of 'chatbots', which were simpler systems outside academic research but which often performed well. There were significant bodies of work on Speech Processing and on MT. Related to these work had started to develop on machine learning (ML) or statistical methods (SM), and formal evaluations of NLP systems had started to be undertaken.

This section reviews each of these topics briefly.

No conclusions about the relative importance of these individual themes should be drawn from the length of coverage of these topics in this section. Nor should any conclusion be drawn about the scale of work (in terms, e.g., of papers or research projects) in the mid-1990s or at other times from the length of coverage here. The level of detail here is driven in part by the expertise of the authors, and in part in order to establish a foundation for our review of the current state of the art in later sections. In particular, the relatively lengthy coverage we give of Jelinek's MT work is because of its role in creating the current focus on data-driven work and SM in the field.

2.1 Parsing

The first strand had a long history back to the IBM parsing systems of the 1970s (based on earlier work by Rosenbaum and Lochak (1966)) which tended to produce very large numbers of possible parsings between which there was no way of choosing. Research methods were changed by the availability of the Penn Treeback (PTB) in 1992 (Marcus, Marcinkiewicz and Santorini 1993) which enabled two things: first, the possibility of learning the rules that had produced the hand parses of the PTB, and secondly, of shifting the goals of the parsing task so that it became necessary only to provide chunks of parsed text across sentence spans and not necessarily a parsing reaching the 'magic' S (sentence) node at the top of the tree.

Thompson (1983) had introduced a taxonomy of work in Computational Linguistics and NLP dividing it into Computation in service to Linguistics; Computation in service to Psychology; Computational Linguistics; Computational Psycholinguistics; Theory of Linguistic Computation and Applied Computational Linguistics. Work in NLE proper, of course, falls in the last category.

Several items on sentence parsing appeared in Volume 1 of *NLE*, including Han and Choi (1995) and Brown (1995), despite the lack of evidence that these sorts of systems had practical applications. Within Thomson's scheme they are probably best classified as Computation in service to Linguistics where the object is to test and develop linguistic theory (Thompson 1983).

2.2 Information extraction

From about 1987, the US funding agency Defense Research Projects Agency (DARPA) effectively created a new Natural Language technology with its MUC whose task was to extract 'messages' of content from often ill-formed and truncated military messages, such as those between ships. Since these were often ungrammatical, there was no hope that a syntactic parser would parse them, and so a kind of surface skimming of input was devised to grasp what was being said, as opposed to a deep, slow, grammar analysis. That technology was called IE and one of its roots

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

was in earlier work at Lancaster University in the 1960s and a program called CLAWS4 which was designed to do automatic part-of-speech tagging: the first program systematically to add to a text 'what it meant' even at a low level (Garside 1987). IE built on that to locate names in text, along with their semantic types, and relate them together by structures called *templates* into what we could call 'forms of fact', which have the form subject–verb–object (e.g. 'John took a job'), but which were also, as we noted above, structured database entries. Other sources of this work include the very early and barely implemented (Wilks 1967/2007) which indicated the feasibility of using semantic preferences for this sort of work. Also important was the much later FRUMP system which demonstrated the feasibility and effectiveness of text skimming technology (De Jong 1982).

By the mid-1990s, IE technology had become well established and was starting to focus on engineering large-scale and deployable applications. Volume 1 of *NLE* includes two good representative papers of the state of the art at the time: Evans et al. (1995)[a] and Friedman, Hripcsak, DuMouchel, Johnson and Claytonal (1995).

2.3 Dialogue and chatbots

In the field of modelling human conversational dialogue with machines, there has been something of the same opposition we saw between deep but ineffectual syntactic parsers and nimble, shallow but effective methods like IE. In the latter part of the last century, there were a number of established systems based on Artificial Intelligence (AI) or linguistics that computed over belief structures (e.g. Ballim and Wilks 1991/2018), conversational models or pragmatic speech acts but none capable of carrying out lengthy conversations with people. But there had been early methods to create dialogue with machines, of which ELIZA in the late 1960s is the best known (Weizenbaum 1966), but a far better program was PARRY, from a real psychiatrist, Ken Colby at Stanford, which mimicked a paranoid patient in a military hospital (Colby 1973). It conducted thousands of hours of dialogue on the early version of the Internet. Like IE, PARRY had no grammar, parsing or logic but only a table of some six thousand patterns that were matched onto its input.

PARRY was very close in style to the movement later called 'chatbots', which was started outside AI by amateur programmers: generally with simple systems that carried on a conversation by means of tricks and ways of not quite understanding what was said, much as many people seem to. In the early 1990s despite a great deal of work, these systems had progressed little beyond PARRY in terms of their ability to convincingly mimic human levels of performance (as demonstrated by early responses to the Loebner prize (Floridi, Taddeo, and Turilli 2009), and indeed one of the present authors (YW) led a team that won the prize in 1997 with a program that fooled at least one judge). However, by the mid-1990s they were more widely available, in large part due to much freer access to computer power by that period compared to the 1960s and 1970s, and many chatbots had developed into commercial applications as front ends to travel reservation and customer service sites.

One such early chatbot system was described in Volume 1 of *NLE* – the PLUS system (Prince and Pernel 1995). This was a system intended to deal with (text) dialogues like:

User: I need a car to go to the airport.
Bot: Do you want a taxi company?
User: No I prefer to rent something.
…

In other words PLUS was an early attempt to deal with the sort of dialogue which is now commonplace with Alexa and its competitors.

[a]Both the current authors were involved, if only peripherally, with the POETIC project and its predecessor TIC from which the Evans et al. (1995) paper arose.

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

One of the striking things rereading the Prince and Pernel paper to prepare this article was the paucity of example dialogues compared to Colby paper published over 20 years previously, illustrating the lack of progress in the intervening period. The work of Prince and Pernel was based on the belief that the limitations of PARRY could only be overcome by the use of deep processing, involving sophisticated models of dialogue, syntactic processing and world knowledge. Applying these deeper techniques was considerably more computationally demanding than the shallow pattern matching techniques used by Colby in PARRY, and used by other later chatbots. Large-scale computational resources were still rare and expensive in the early 1990s so perhaps the lack of real progress is unsurprising.

2.4 Speech and MT

It had always been the intention that *NLE* should include papers on MT and Speech, and the original aims and scope statement in the Volume 1 Issue should perhaps have made this more explicit. There were very few such papers in the first volume: one on speech synthesis, mainly focused on the analysis of incoming text (Bachenko, Fitzpatrick and Daugherty 1995), and none on MT *per se*, although two papers had mathematical or technical relevance to both MT and speech analysis (MacKay and Bauman Peto (1995) and Church and Gale (1995)). The first paper to appear in *NLE* directly relevant to MT was Cranias, Papageorgiou and Piperidis (1997).

Since the 1980s effective single-speaker, limited vocabulary, speech recognition systems had been available, and by 1995 steady progress was being made in terms of vocabulary size, training time for individual speakers, multi-speaker operation, noise tolerance and so on. Hidden Markov Models had proved an effective technique to deal with the variability in the speech signal in practice and had been widely adopted (Juang and Rabiner 2005). Large vocabulary speech synthesis systems were available, but had an unnatural 'robotic' quality, and were sometimes difficult to follow, certainly if used for any length of time.

By the mid-1990s, Translation Memory (Nagao 1984) was a well-developed technology, in use by increasing numbers of human translators, but still the subject of continuing research (Cranias *et al.* 1997; Somers 2003). However, fully automated MT was at a rather odd juncture. After 35 years of research and development, there were MT Systems in practical use in a variety of contexts. SYSTRAN was an old but reliable system, rewritten and patched up over the previous 25 years for Russian-to-English. It had also been in use since the 1970s by the European Commission in Luxembourg to produce rough drafts of hundreds of memoranda a day. It provided reasonable translations at about the 65% correct level (of sentences translated) over a wider range of language pairs that usually included English (Hutchins and Somers 1992).

Despite substantial expenditure of time and effort on research and development, it had proved hard to exceed these figures using any of the then current techniques. In the 1980s, the European Commission devoted over 75 million euros (ecus at the time) to a European rival to SYSTRAN (EUROTRA) that succeeded in producing only a handful of translated sentences (Hutchins and Somers 1992; Oakley 1993).

In the late 1980s, a research group at IBM New York, led by Fred Jelinek, a distinguished speech scientist and engineer, announced that they were beginning a program of MT research and development, one quite unlike any previous work in that field.

Jelinek and his colleagues had designed a translation system called CANDIDE that learned how to translate from English to French, and vice versa, simply from processing hundreds of millions of words of French–English parallel text, which is the form in which Hansard, the record of proceedings of the Canadian Parliament, is published. The choice of the name CANDIDE, from Voltaire's novel of that name, was not accidental: Candide was famously young and fresh, and uncontaminated by the prejudices of the then learned world. Similarly, CANDIDE did not start out prejudiced by the grammar or dictionaries or even any knowledge of the French language!

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

Jelinek famously boasted that CANDIDE got better whenever he fired a linguist from his team, a joke that was making the same cultural point we noticed earlier: it was a project of speech engineers, designed to extend the success of speech recognition methods into the logic and linguistics-dominated area of translation. CANDIDE's core algorithm defined a statistical computation over the Hansard data that captured the regularities in it and allowed the processing of more, previously unseen, text.

The details of how Jelinek's system actually worked need not detain us: the important point is that it was a series of comparisons of millions of words of French and English, and that a statistical value was found such that it selected the best equivalent in the other language for any new sentence it had not seen before. One of those comparisons was, for example, to seek out which French words appeared most in the sentences that had been aligned one-to-one with English sentences containing a particular English word. The parts of CANDIDE were reassembled by what Jelinek called a general equation of MT, which was a very large statistical expression whose maximum value was to be found by training so as to give the best possible translation for a new segment of text. This idea of a machine being 'trained' to translate, by seeing many examples of translations, is an example of what is now generally called ML. In fairness, too, it should be pointed out that some language scientists, in particular Nagao in Japan, had already claimed that it should be possible for a computer to learn to translate from a large number of actual translations, and that suggestion owed nothing to earlier speech research.

Many were astonished that CANDIDE did as well as it did, which was to get about half the sentences given to it more or less correctly translated. Given that it knew no French grammar or vocabulary when it started but learned everything just from millions of words of actual translations, it was a remarkable result, and undoubtedly said something about the meaning content transferred in translation and its relationship to SM.

By the mid-1990s, Jelinek had begun to develop techniques for building dictionaries and grammar rules of the kinds used by linguists but he did so not by intuition, which is to say by a linguist writing those structures down for a computer to use. Instead, he developed techniques, again based on speech engineering, so that a computer could learn dictionaries and grammar rules directly from large corpora of text (Chelba and Jelinek 2000). He thus renounced his distaste for what we could call linguistic structures, but insisted that the structures be based on data, and learned from actual texts: sometimes from one language (for grammar rules) and sometimes from two like Hansard (for dictionaries).

That work began a revolution in NLE which continues to this day. This revolution is based on the adoption of ML and/or SM, an approach which had been neglected in the field between the 1960s and 1980s.

In Volume 1 of *NLE*, there were already two papers firmly based within the SM tradition: Church and Gale (1995) and MacKay and Bauman Peto (1995); and a third which perhaps is closer to modern ML: Han and Choi (1995).

2.5 Formal evaluations, ML and SM

In the USA, and to some extent worldwide, the field of NLE has been advanced by the open competitions fostered by the major US funding agencies such as DARPA. For US researchers, participation in such competitions has become a condition of being funded, but the competitions are open to teams from anywhere. They have been over a wide range of linguistic phenomena from part-of-speech tagging right through to pragmatic and logical inference. As previously mentioned, they began with the analysis of ill-formed naval messages (the MUC, Message Understanding Competition, in 1987, see Grishman and Sundheim 1996 *op cit*) based on shared data, and later of components between teams, and have undoubtedly seen monotonic changes in scores, usually based not only on very complex blind scoring systems, but also relying to some extent on the honesty of research teams.

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

By the mid-1990s, this model of evaluation had started to become a central part of the field, stimulated in part by Sparck Jones and Galliers (1995), as well as the development of more voluntary activity (e.g. Palmer and Finin 1990) to parallel the US research funders' programs. This development had led to such things as much larger test collections and sets of available useful resources.

These larger test collections and sets of useful shared resources were the underpinning of the move to much larger scale applications, and the broader use of ML and SM in the field.

However, this change did not immediately become apparent in the material published in *NLE*. Some papers in Volume 1 were based on extremely small-scale experiments. For example, Brown (1995) describes tests on corpora of 129 and 100 sentences, little better than Winograd's thesis published 20 years ago (Winograd 1973). Others were starting to use more substantial data sets. For example, MacKay and Bauman Peto (1995) based their conclusions on a collection of over 2 million words (substantial for its day).

2.6 Concluding remarks on the 1990s

We begin by presenting a tabular analysis of the papers which appeared in Volume 1 of *NLE* to provide a somewhat anecdotal overview of the state of the art of the field in 1995.

Several papers were excluded from the analysis as either being survey papers on particular topics or not being amenable to this analysis. It is included here principally for completeness and to provide a basis from comparison in the discussion of the current state of the art in Section 3.

In contrast to Table 2, representing the current state of the art of NLE, Table 1 shows that in 1995 almost all papers published by *NLE* relied on the use of larger linguistic structures or information: sentences, clause, phrases or at the very least words with parts-of-speech assignment were the product of the analysis.

In conclusion, in 1995, there were the first signs of the ML revolution appearing in NLE, especially in MT. A more scientific model of research, dominated by relatively large shared test data sets and other resources had started to emerge. But much work was still based on small-scale handcrafted grammars, software and other resources, which was little closer to practical application with real users than the systems had been around 20 and more years previously, in the 1970s.

3. The state of the art 25 years later

In contrast to the state of the art of NLE in the early 1990s, we are now at the end of the second decade of the 21st century and want to pick out three new, or at least much more prominent, aspects of the field that were not so prominent 25 years ago.

First is the widespread, practical use of NLE products by millions, possibly billions, of people worldwide. We note, in particular, the very widespread use of MT in Facebook, Google and elsewhere. Much of the criticism of the ALPAC report in the 1960s (Pierce et al. 1966) remains valid. These systems cannot produce the quality of translations produced by highly skilled human translators on a consistent basis. They often produce gobbledygook, especially for languages where they have been less well trained and for things like fast-moving street slang on social media, and other specialist languages, but few human translators can produce good translations in these circumstances. However, fully automatic MT systems do provide sufficient quality for many practical tasks: to assess the general tone and drift of reaction to a social media post; to help assess the relevance of a patent for human translation and technical analysis; to access the content of a scientific article where the reader's expert knowledge compensates for translation errors; and so on. Furthermore, there is evidence that the performance of MT systems is sometimes better than available human translations in these practical settings, a circumstance the authors of the ALPAC report foresaw as a possibility, despite their focus on scientific and technical translation in

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

Table 1. Shallow vs Deep techniques of 20 recent *NLE* articles

Paper		BoW/n-gram/skip gram[a]	Chunking/POS	Clause/Sentence Parsing
1.	Langlois et al. (2018)	X		
2.	Fatima et al. (2018)	X		
3.	Derici et al. (2018)		X	X
4.	Hirano et al. (2018)		X	X?
5.	Wei et al. (2018)	X		
6.	Li et al. (2018)	X		
7.	Braun, Reiter and Siddharthanl[b] (2018)			X
8.	Banea and Mihalcea (2018)	X	X	X
9.	Laddha and Mukherjee (2018)	X		
10.	Perinan-Pascual (2018)	X		
11.	Chen et al. (2018)	X		
12.	Gründer-Fahrer et al. (2018)	X		
13.	Biemann, Faralli, Panchenko and Ponzetto (2018)	X		
14.	Kübler et al. (2018)	X	X	
15.	Marrero and Urbano (2018)	X	X	
16.	Kadari et al. (2018)	X		
17.	Garcia, Gómez-Rodríguez and Alonso (2018)			X
18.	Azmi and AlShenaifi (2017)			X
19.	Giannella et al. (2017)	X	X	
20.	Krüger et al. (2017)	X	X	

[a]Includes tokenisation, stemming, etc.; [b]generation.

Table 2. Shallow vs Deep techniques of *NLE* Volume 1 articles

Paper		BoW/n-gram/skip gram[a]	Chunking/POS	Clause/Sentence Parsing
1.	Justeson and Katz (1995)	X	X?	
2.	Friedman et al. (1995)			X
3.	Prince and Pernel (1995)			X
4.	Han and Choi (1995)			X
5.	Bachenko, Fitzpatrick and Daugherty (1995)		X[b]	
6.	Pulman (1995)			X
7.	Mikheev and Liubushkina (1995)		X	
8.	Wintner and Ornan (1995)			X
9.	MacKay and Bauman Peto (1995)	X		
10.	Brown (1995)			X
11.	Evans et al. (1995)		X	X

[a]Includes tokenisation, stemming, etc.; [b]actually phrase parsing.

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

a very different world. However, this possibility probably leads to the redundancy of such human translators, so the evidence is hard to gather (see Läubli and Orrego-Carmona 2017).

Further, many operational systems in search, sentiment analysis, and a number of other areas rely on technologies which would have been regarded as topics for NLE research 25 years ago, but are now in widespread use (although most if not all remain topics for active research). Probably most notable is the widespread use of named entity recognition and tracking.

Another is chatbot technology. By the mid-2000s, this technology had started to became mainstream via the annual Loebner competition (Floridi *et al.* 2009) and is now the inspiration behind modern applications such as Siri and Alexa. Some commentators believe there will be as many copies of these as there are humans within decades. The vast majority of users simply take these commercial products for granted, whereas those of us who have some concept of the detailed operations and the developments over the last quarter century or so which have led to them are often left amazed by the progress and the successes of our peers.

Second, as in many other fields, the current dominance of ML is noteworthy. Supervised, semi-supervised and unsupervised ML usually over very large corpora has become the dominant paradigm. Some thinkers always expected this to be the case. Sparck Jones (1986) discussing work she originally undertook in the 1960s is a notable example. Although the terminology used is very different from current usage: unsupervised learning is referred variously by her as 'the Theory of Clumps' or 'automatic classification', yet the relationship between the ideas is clear (Wilks and Tait 2005). As an illustration of this point, all but one paper in *NLE* 24(5) leant significantly on ML techniques, and even that paper might be argued to make some use of ML.

Recently, Deep Learning has become a widely adopted within the NLP community. Deep Learning is a coverall term for a range of SM and ML methods which are characterised by the use of hidden layers or other stack-like arrangements of individual learning systems. Generally, these hidden layers are thought to capture some sort of generalisation or abstraction not easily captured by simpler ML architectures, and thereby to address sparse data issues amongst other things. Significant successes have been reported on a number of long-standing problems (e.g. in parsing, Andor et al. 2016), but there remain many problems to be overcome before the true impact of Deep Learning on NLE can be assessed (Manning 2015; Cho 2018).

Third, the widespread use of word n-gram, skip-gram and related models at the expense of deeper structures is striking. Much work in the early 1990s was based on the assumption that successful applications required accurate and robust analysis based on the grammatical structures in natural language utterances. Jelinek's insight – which was Colby's for conversational analysis two decades earlier – that deeper structural analysis might not be essential for useful MT has proved applicable to a range of other applications. Though Jelinek, as we noted, eventually accepted that his MT systems needed to learn grammatical and lexical information beyond word-string-to-word-string matching. It seems that learned pattern matching across strings of words (or morphologically reduced word lemmas) is adequate for many practical purposes. A similar movement has taken place in modelling conversation: whereas chatbot systems were for decades excluded from NLP research as superficial and skimming only surface phenomena in input, they have now developed to a point of sophistication where they are both in widespread commercial use – in web-based help systems – and are being melded to 'deeper' analysis systems that attempt to understand more fully what is being said to them.

To illustrate this move from attempting to extract deeper structures from text to the use of shallower structures, we analysed 20 recent articles published in *NLE*. The survey was undertaken in late 2018. Twenty papers were selected from *NLE* Volumes 24 (up to Issue 5) plus 23(6), and to make up the numbers, the first paper from 23(5). Volume 24(3) was omitted from the analysis. It was a special issue on Language for Images, and the Special Issue selection policy meant it was an even less representative sample of the field than regular issues. Also excluded were survey articles (which often cover older as well as more recent work); papers on text generation and papers which focus on specific topics, like tokenisation methods rather than end-to-end applications.

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

So from this limited survey, 75% of papers published in the relevant issues of *NLE* use bag of words, n-grams or skip grams. About 50% of the papers use those methods alone. About 35% use chunking (into phrases, clauses or in some cases less linguistically motivated units) or part-of-speech tagging. About 30% of the published papers aim to produce structures at the clause or sentence level. About 35% of the papers use a combination of techniques.

A striking feature is the variety of languages covered in these recent papers. They include Turkish, Arabic, Yoruba, Chinese and a variety of Romance languages including less widely spoken ones like Galician. In contrast, in *NLE* Volume 1, all but three of the papers concerned English (exclusively or very predominantly). The exceptions addressed French, Swedish (both in a paper substantially focused on English) Russian and Hebrew. However, note that many fewer papers were published in Volume 1 than in Volume 24: 14 in Volume 1 as opposed to 32 in Volume 24.

One might speculate about why these techniques based on bags of words, n-grams or skip grams are proving so popular in practically orientated work. One possibility is that much of the meaning of an utterance is carried in the lexical elements, and that linguists of an earlier era overestimated the extent to which syntax, structural semantics and pragmatics contributed to the use of language in practical settings. Another possibility is that language is inherently redundant, and in practice, information in the lexical channel reinforces and supplements information in the deeper structural channels (e.g. relying on grammatical relations), so the structure as opposed to the words themselves can often be ignored. A third is that current applications of NLE are biased towards those in which these n-gram-based techniques work well. A fourth is that the development of techniques which utilise the shallow technologies has been stimulated by the structure of the field at the expense of other directions.

There is no doubt that open competitions, based on a shared task, with fairly clear evaluation metrics, shared data and often other shared resources, have advanced the field; the doubts some have are mainly as to whether the isolated, testable tasks add together to give advance in more holistic applications like MT or conversational analysis. European research funders have tended to shy away from such competitions, seeing them as a waste of resources. But even after over 30 years, the model pioneered by MUC continues to have an important influence on research in the field.

In particular, the shared tasks represented by MUC and its successors have led to the easy availability of freely available data sets which are essential for the development of ML-based systems.

However, as well as the limitations of isolated, testable tasks, as opposed to end-to-end applications, another doubt about the open competitions is whether the metrics used to evaluate success have been biasing technological development in ways which inhibit innovation. For example, the BLEU metric (Papenini, Rouskos, Ward and Whu 2002), widely used in MT evaluation and elsewhere, operates by comparing the output of a test system with a gold standard example text on an n-gram by n-gram basis. One has to question whether continued widespread use of BLEU as a measure of MT system quality inhibits work which may translate using paraphrase or radical rewording, compared to work which translates word by word and n-gram by n-gram, especially in the light of doubts about its correlation with human assessment of MT quality (Callison-Burch, Osborne and Koehn 2006).

Nonetheless, the probability is that the move away from deeper analysis has been driven more by the nature of language used in practical day-to-day settings than any other factor.

4. So what has changed?

As noted in the previous section, the widespread use of word (or lemma) n-grams and skip grams is still a surprise to many established researchers. Although this trend had started to emerge in the 1980s outside the information retrieval (IR) community, it was not widely adopted by the

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

NLE community until the first decade of the 21st century, stimulated in part by interactions and crossovers between work in the IR and NLP communities.

In the 1960s and 1970s, there was an ongoing debate within IR as to whether document retrieval systems using statistics derived from word usage in documents and cross-document relationships (like citations) could match or exceed the performance of knowledge-based system using controlled vocabularies, taxonomic indexing schemes and other hand-crafted knowledge sources, familiar in related but rather different forms to the library science and NLP communities. By the 1990s, it had been convincingly demonstrated that practical IR systems (like internet search engines) could be constructed using primarily word-based statistics, and this evidence had started to excite interest in a number of areas, especially ones of interest to both the IR and NL communities, like MT and word sense disambiguation.

Indeed, it was often the case that systems based on word grams or skip grams were seen initially as baseline systems against which 'smarter' systems could be compared. It just proved harder to beat the baseline than the authors expected. The direct influence on MT developments come from the automatic speech recognition community through Jelinek, as we noted, but indirect influence on NLP and AI came at the same time from the IR community through researchers like Sparck Jones and others.

The field has become significantly more scientific, in the sense that it is much more dominated by evaluation regimes in which a number of systems are measured against some agreed benchmark, and that advance was driven almost entirely by the US funding agencies. This has encouraged better formulated hypotheses about ways to improve systems and performance and more-or-less objective assessment of whether performance has in fact been improved. As noted previously, one can have various doubts about the effect of this change to a more scientific methodology has had in the development of overall system performance and for some specific technologies. However, published papers in *NLE* are now, at the time of writing in 2018, much more likely to report facts expressed as proof or disproof of hypothesis which can be reproduced, than was the case in the mid-1990s.

Formal evaluation campaigns have on the whole improved dialogue in the field. If one goes back beyond the genesis of *NLE* to the 1970s when both the current authors were already in the field, much published material went little beyond what might be labelled feasibility studies (see, e.g., Schank and Colby 1973). This style, and the poorly engineered nature of the systems produced, combined with the limitations of the then available computer power, and the poor availability of large-scale data, inhibited progress in the field for many years.

Today, the widespread availability of open source tools of a wide variety of kinds – a product very largely of the insistence on group cooperation by the US funding agencies – makes it possible to rapidly bolt together end-to-end applications and to produce specific variations of systems and to observe the effect on performance of changing them.

Of course, true reproducibility remains an elusive goal as in many other fields. But great progress has been made over the last 25 years.

5. What has not progressed as we might have expected?

In the 1990s, there was a general expectation that much future language work would be based on widespread use of pre-existing dictionaries. Pioneering work in this area had been undertaken by Michiels in Liege in the 1980s (Michiels 1983) and this was followed by Boguraev, Carroll, Briscoe, Carter and Grover (1987), Boguraev and Briscoe (1989), Wilks, Slator and Guthrie (1996), and many others. Much of the work was based on the *Longmans Dictionary of Contemporary English* (Proctor 1978), although a number of other dictionaries were used. There was an untested assumption that dictionaries were where meaning was located so let's go and use them. So it was widely expected that the future of the field lay with the use and development of these dictionaries, and semantic data extracted from them. However, this has proved not to be the case.

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

In practice, the dominant lexical resource has been Wordnet[b] and the related projects for other languages (Fellbaum and Miller 1998; Bond and Paik 2012), which is paradoxical given that Wordnet was designed as a psychological rather than a computational tool.

One reason for the eventual dominance of Wordnet was the legal problems created by the commercial publishers who felt they needed to control the intellectual property embodied in the machine-readable version of their dictionaries, and the consequent problems with licensing and reuse this created. Wordnet specifically allows such usage.

Another reason commercial dictionaries were not as widely used in NLE as might have been expected was the somewhat divergent interests of the natural language engineers and the lexicographers when the dictionary needed to be updated and extended.

A related area in which progress has not been made in the direction generally expected in the early 1990s is the use of existing reference works. It had been expected that NLE systems in the 21st century would depend significantly on what one would now call fact bases mined from standard references like the *Encyclopaedia Britannica.* This has not happened in part because, just as with dictionaries, such publications have been supplanted to a large extent in general use by internet resources like Wikipedia. Resources that can be updated, verified and used by anyone have won out over closely controlled 'walled-in' resources. Despite this, our ability to ingest these open human usable resources into a computer system and use them effectively for purposes like question answering has progressed to a much more limited extent than many expected in the mid-1990s and it remains an active area of research (see, e.g., Choi, Seo, Chen, Jia and Berant 2018). However, Google's Knowledge Graphs[c] , derived from Wikipedia, have actually proved a useful source of real-world knowledge for a range of applications including their commercial dialogue system SIRI.

It has proved much more challenging to deal with notions of truth or veracity in practical language than most language engineers envisaged 25 years ago. It was simply not envisaged in the mid-1990s that determining whether a newspaper article was reporting facts or merely repeating fake news would became a major future task of NLE systems.

Finally, there has been much less progress in some important and pervasive phenomena of real language use than one might have expected. For example, perhaps most notably, there has been less progress with metaphor processing than some would have hoped for 25 years ago. After a gap of several years in 2011, the US funding agency iARPA[d] did restart major funding of metaphor research at a number of sites and some important work emerged in a number of languages (English, Spanish, Russian and Farsi). Cross linguistic comparisons were made and benchmarked and some useful progress was demonstrated. However, no attempt was made to link the work directly to MT or other practical language processing tasks, so metaphor remains a pervasive and central unsolved issue of MT and language engineering generally.

6. Conclusions

Much has changed in NLE since Volume 1 of *NLE* was published in 1995. The field is much more scientific, works with much larger data sets and addresses more ambitious tasks than was the case 25 years ago.

Above all, language engineering technologies are now in day-to-day use by millions of ordinary people around the world. Examples include the translate button on Facebook posts and dialogues with Alexa. As recently as the 1990s, such systems were the subject of very limited applications in specialised areas, tiny feasibility studies or were little more than pipe dreams in research labs.

[b] See https://wordnet.princeton.edu/ (checked 15 October 2018).
[c] See https://en.wikipedia.org/wiki/Knowledge_Graph (checked 26 January 2019)
[d] See https://www.iarpa.gov/index.php/research-programs/metaphor/baa (checked 26 January 2019)

https://doi.org/10.1017/S1351324919000081 Published online by Cambridge University Press

Furthermore, there is extremely widespread use of NLE technologies like named entity recognition, IE, sentiment analysis, speech recognition and many more, as components of widely used information systems like search engines and customer service systems.

The field has grown enormously in terms of engineers and scientists involved, languages covered, papers produced and so on. NLP or NLE is now a widely accepted and well-known discipline with a workforce in commercial demand in a way that was not the case in the recent past.

Much progress has been made, but there remain many challenges. Some essentially technical challenges were noted in the previous section, and technology has been inevitably the focus of this piece. However, it would be wrong to close without mentioning the biggest challenge for the next 25 years of *NLE*: the ethical implications of pervasive technologies which can in some real sense understand human language.

References

Andor D. Alberti C., Weiss D., Severyn A., Presta A., Ganchev K., Petrov S. and Collins M. (2016). Globally Normalized Transition-Based Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany. 2442–2452.

Azmi A. and Alshenaifi N. (2017). Lemaza: An Arabic why-question answering system. *Natural Language Engineering* 23(6), 877–903. doi:10.1017/S1351324917000304

Bachenko J., Fitzpatrick E. and Daugherty J. (1995). A rule-based phrase parser for real-time text-to-speech synthesis. *Natural Language Engineering* **1**(2), 191–212. doi:10.1017/S1351324900000140

Ballim A. and Wilks Y. (1991/2018). *Artificial Believers: The Ascription of Belief*. New Jersey: Ablex Books; reprinted by Routledge, London.

Banea C. and Mihalcea R. (2018). Possession identification in text. *Natural Language Engineering* **24**(4), 589–610. doi:10.1017/S1351324918000062

Biemann C., Faralli S., Panchenko A. and Ponzetto S. (2018). A framework for enriching lexical semantic resources with distributional semantics. *Natural Language Engineering* **24**(2), 265–312. doi:10.1017/S135132491700047X

Boguraev B. and Briscoe T. (Eds) (1989). *Computional Lexicography for Natural Language Processing*. Harlow, Essex, England: Longman.

Boguraev B.K., Garigliano R. and Tait J.I. (1995). Editorial. *Natural Language Engineering* **1**(1), 1–7.

Boguraev B., Carroll J., Briscoe E., Carter D. and Grover C. (1987). The Derivation of a Grammatically-Indexed Lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, CA. 193–200.

Bond F. and Paik K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71

Braun D., Reiter E. and Siddharthan A. (2018). SaferDrive: An NLG-based behaviour change support system for drivers. *Natural Language Engineering* **24**(4), 551–588. doi:10.1017/S1351324918000050

Brown J.C. (1995). High speed feature unification and parsing. *Natural Language Engineering* **1**(4), 309–338.

Callison-Burch C., Osborne M., Koehn P. (2006). Re-evaluation the Role of Bleu in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*. Trento, Italy. 249–256.

Chelba C. and Jelinek F. (2000) Structured language modeling. *Computer Speech & Language* **14**(4), 283–332. doi:10.1006/csla.2000.0147

Chen Y., Zheng Q., Tian F., Liu H., Hao Y. and Shah N. (2018). Exploring open information via event network. *Natural Language Engineering* **24**(2), 199–220. doi:10.1017/S1351324917000390

Cho K. (2018). Deep learning. In Mitkov R. (ed), *The Oxford Handbook of Computational Linguistics*, 2nd Edition. Oxford, England: Oxford University Press. doi:10.1093/oxfordhb/9780199573691.013.55

Choi E., Seo M., Chen D., Jia R. and Berant J. (2018). *Proceedings of the Workshop on Machine Reading for Question Answering*. Melbourne, Australia: Association for Computational Linguistics.

Church K.W. and Gale W.A. (1995). Poisson mixtures. *Natural Language Engineering* **1**(4), 163–190.

Colby, K.M. (1973). Simulation of Belief Systems. In Schank R.C. and Colby K.M. (eds), *Computer Models of Thought and Language*. San Francisco: W.H. Freeman and Co. 251–286.

Cranias L., Papageorgiou H. and Piperidis S. (1997). Example retrieval from a translation memory. *Natural Language Engineering* **3**(4), 255–277

Cunningham H. (1999). A definition and short history of language engineering. *Natural Language Engineering* **5**(1), 1–16.

De Jong G.F. (1982). An overview of the FRUMP system. In Lehnert W.G. and Ringle M.H. (eds), *Strategies for Natural Language Processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Derici C., Aydin Y., Yenialaca Ç, Aydin N., Kartal G., Özgür A. and Güngör T. (2018). A closed-domain question answering framework using reliable resources to assist students. *Natural Language Engineering* **24**(5), 725–762. doi:10.1017/S1351324918000141

Evans R., Gaizauskas R., Cahill L.J., Walker J., Richardson J. and Dixon A. (1995). POETIC: A system for gathering and disseminating traffic information. *Natural Language Engineering* **1**(4), 363–387.

Fatima M., Anwar S., Naveed A., Arshad W., Nawab R., Iqbal M. and Masood A. (2018). Multilingual SMS-based author profiling: Data and methods. *Natural Language Engineering* **24**(5), 695–724. doi:10.1017/S1351324918000244

Fellbaum C. and Miller G.A. (1998). *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Floridi L., Taddeo M. and Turilli M. (2009). Turing's imitation game: Still an impossible challenge for all machines and some judges—an evaluation of the 2008 Loebner contest. *Minds & Machines* (19):145–150. doi:10.1007/s11023-008-9130-6.

Friedman C., Hripcsak G., DuMouchel W., Johnson S.B. and Clayton P.D. (1995). Natural language processing in an operational clinical information system. *Natural Language Engineering* **1**(1), 83–108.

Garcia M., Gómez-Rodríguez C. and Alonso M. (2018). New treebank or repurposed? On the feasibility of cross-lingual parsing of Romance languages with Universal dependencies. *Natural Language Engineering* **24**(1), 91–122. doi:10.1017/S1351324917000377

Garside R. (1987). The CLAWS Word-tagging System. In Garside R., Leech G. and Sampson G. (eds), *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Giannella C., Winder R. and Petersen S. (2017). Dropped personal pronoun recovery in Chinese SMS. *Natural Language Engineering* **23**(6), 905–927. doi:10.1017/S1351324917000158

Grishman R. and Sundheim B. (1996). Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING), I*, Copenhagen, 466–471.

Gründer-Fahrer S., Schlaf A., Wiedemann G. and Heyer G. (2018). Topics and topical phases in German social media communication during a disaster. *Natural Language Engineering* **24**(2), 221–264. doi:10.1017/S1351324918000025

Han Y.S. and Choi K.-S. (1995). Best parse parsing with Earley's and Inside algorithms on probabilistic RTN. *Natural Language Engineering* **1**(2), 147–161.

Hirano D., Tanaka-Ishii K. and Finch A. (2018). Extraction of templates from phrases using Sequence Binary Decision Diagrams. *Natural Language Engineering* **24**(5), 763–795. doi:10.1017/S1351324918000268

Hutchins J. and Somers H. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Juang B.H. and Rabiner L.R. (2005). *Automatic Speech Recognition– A Brief History of the Technology Development*. Georgia Institute of Technology, Atlanta. https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf (Checked 10 December 2018)

Justeson J. and Katz S. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* **1**(1), 9–27. doi:10.1017/S1351324900000048

Kadari R., Zhang Y., Zhang W. and Liu T. (2018). CCG supertagging with bidirectional long short-term memory networks. *Natural Language Engineering* **24**(1), 77–90. doi:10.1017/S1351324917000250

Krüger K., Lukowiak A., Sonntag J., Warzecha S. and Stede M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering* **23**(5), 687–707. doi:10.1017/S1351324917000043

Kübler S., Liu C. and Sayyed Z. (2018). To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. *Natural Language Engineering* **24**(1), 3–37. doi:10.1017/S1351324917000298

Laddha A. and Mukherjee A. (2018). Aspect opinion expression and rating prediction via LDA-CRF hybrid. *Natural Language Engineering* **24**(4), 611–639. doi:10.1017/S135132491800013X

Langlois D., Saad M. and Smaliki K. (2018). Alignment of comparable documents: Comparison of similarity measures on French–English–Arabic data. *Natural Language Engineering* **24**(5), 677–694. doi:10.1017/S1351324918000232

Läubli S. and Orrego-Carmona D. (2017). When Google Translate is better than Some Human Colleagues, those People are no longer Colleagues. In *Proceedings of Translation and the Computer 39, Asling, the International Association for Advancement in Language Technology*, London. 59–69.

Li B., Gaussier E. and Yang D. (2018). Measuring bilingual corpus comparability. *Natural Language Engineering* **24**(4), 523–549. doi:10.1017/S1351324917000481

MacKay D.J.C. and Bauman Peto L.C. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering* **1**(3), 289–307.

Manning C.D. (2015). Computational linguistics and deep learning. *Computational Linguistics* **41**(4), 701–707.

Marcus M.P., Marcinkiewicz M.A. and Santorini B. (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* **19**(2), 313–330.

Marrero M. and Urbano J. (2018). A semi-automatic and low-cost method to learn patterns for named entity recognition. *Natural Language Engineering* **24**(1), 39–75. doi:10.1017/S135132491700016X

Michiels A. (1983). Automatic analysis of texts. In Jones K.P. (ed), *Informatics 7: Intelligent Information Retrieval*. Cambridge: Aslib, pp. 103–120.

Mikheev A. and Liubushkina L. (1995). Russian morphology: An engineering approach. *Natural Language Engineering* **1**(3), 235–260. doi:10.1017/S135132490000019X

Nagao M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In Elithorn A. and Banerji R. (eds), *Artificial and Human Intelligence. Edited Review Papers Presented at the International NATO Symposium on Artificial and Human Intelligence, 1981*. Lyon, Amsterdam, New York, Oxford, North Holland, pp. 173–180.

Oakley B. (1993). EUROTRA final Review Panel Report. *Commission of the European Communities*. Available from: http://aei.pitt.edu/36888/1/A2903.pdf (Checked 26 January 2019).

Palmer M. and Finin T. (1990). Workshop on the evaluation of natural language processing systems. *Computational Linguistics* **16**(3), 175–181.

Papenini K., Rouskos S., Ward T. and Whu W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia. 311–318.

Periñan-Pascual C. (2018). DEXTER: A workbench for automatic term extraction with specialized corpora. *Natural Language Engineering* **24**(2), 163–198. doi:10.1017/S1351324917000365

Pierce J.R., Carroll J.B., Hamp E.P., Hays D.G., Hockett C.F., Oettinger A.G. and Perlis A. (1966). *Language and Machines — Computers in Translation and Linguistics*. Washington, DC: ALPAC report, National Academy of Sciences, National Research Council.

Prince V. and Pernel D. (1995). Several knowledge models and a blackboard memory for human-machine robust dialogues. *Natural Language Engineering* **1**(20), 113–145.

Proctor P. (ed.) (1978). *Longman Dictionary of Contemporary English*. Harlow, Essex: Longman Group.

Pulman S. (1995). Anaphora and ellipsis in artificial languages. *Natural Language Engineering* **1**(3), 217–234. doi:10.1017/S1351324900000188

Rosenbaum R. and Lochak D. (1966). The IBM core grammar of English. In Lieberman D. (ed), *Specification and Utilization of a Transformational Grammar*. AFCRL-66-270 (1966). Yorktown Heights, New York: Thomas J. Watson Research Center, IBM Corporation.

Schank R.C. and Colby K.M. (Eds.) (1973). *Computer Models of Thought and Language*. San Francisco: W.H. Freeman and Co.

Somers H. (2003). Translation memory. In Somers H. (ed), *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins.

Sparck Jones K. (1986). *Synonymy and Semantic Classification*. Edinburgh: Edinburgh University Press.

Sparck Jones K. and Galliers J.R. (1995). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin: Springer.

Tait J. (2019). Editorial. *Natural Language Engineering* **25**(1), 1–4.

Tait J.I. (ed). (2005). *Charting a New Course: Natural Language Processing and Information Retrieval*. Dordrecht, NL: Springer.

Thompson H. (1983). Natural language processing: A critical analysis of the structure of the field, with some implications for parsing. In Sparck Jones K. and Wilks Y. (eds), *Automatic Natural Language Parsing*. Chichester, England: Ellis Horwood.

Wei Y., Wei J. and Yang Z. (2018). Unsupervised learning of semantic representation for documents with the law of total probability. *Natural Language Engineering* **24**(4), 491–522. doi:10.1017/S1351324917000420

Weizenbaum J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9, 36–45. doi:10.1145/365153.365168.

Wilks Y. (1967). Text searching with templates. Cambridge language research unit, research memorandum. In Ahmad K., Brewster C., Stevenson M. (eds), *Words and Intelligence I. Text, Speech and Language Technology*, vol. 35. Dordrecht: Springer. Reprinted (2007).

Wilks Y.A., Slator B.M. and Guthrie L.M. (1996). Electric Words. Cambridge, Mass: MIT Press.

Wilks Y.A. and Tait J.I. (2005). A retrospective view of synonymy and semantic classification. In *Charting a New Course: Natural Language Processing and Information Retrieval*, pp. 1–11. Springer, Dordrecht.

Winograd T. (1973). A procedural model of language understanding. In Schank R.C. and Colby K.M. (eds), (1973). *Computer Models of Thought and Language*. San Francisco: W.H. Freeman and Co. pp. 152–186.

Wintner S. and Ornan U. (1995). Syntactic analysis of Hebrew sentences. *Natural Language Engineering* **1**(3), 261–288. doi:10.1017/S1351324900000206

Cite this article: Tait J and Wilks Y. Anniversary article: Then and now: 25 years of progress in natural language engineering. *Natural Language Engineering* **25**, 405–418. https://doi.org/10.1017/S1351324919000081