Natural Language Engineering **7** (1): 29–46. Printed in the United Kingdom

© 2001 Cambridge University Press

Automatic language and information processing: rethinking evaluation†

KAREN SPARCK JONES

Computer Laboratory, University of Cambridge,
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK
e-mail: sparckjones@cl.cam.ac.uk

(*Received 12 December 2000*)

Abstract

System evaluation has mattered since research on automatic language and information processing began. However, the (D)ARPA conferences have raised the stakes substantially in requiring and delivering systematic evaluations and in sustaining these through long term programmes; and it has been claimed that this has both significantly raised task performance, as defined by appropriate effectiveness measures, and promoted relevant engineering development. These controlled laboratory evaluations have made very strong assumptions about the task context. The paper examines these assumptions for six task areas, considers their impact on evaluation and performance results, and argues that for current tasks of interest, e.g. summarising, it is now essential to play down the present narrowly-defined performance measures in order to address the task context, and specifically the role of the human participant in the task, so that new measures, of larger value, can be developed and applied.

1 Introduction

The speech and language evaluations sponsored by the US Government in the last two decades have had a major impact on the field. These evaluations, driven by the (Defence) Advanced Research Projects Agency ((D)ARPA) and the National Institute of Standards and Technology (NIST), have been large programmes with many participants, and with series of test cycles over periods of years. They have addressed different tasks, for instance speech recognition and information retrieval, within a common evaluation framework focused on technology assessment using a carefully controlled train-test-measure protocol. However, the apparent similarity between the programmes that this common approach implies conceals significant differences in the task areas. These should be more clearly recognised, so that the inferences to be drawn from the different evaluations about the state of the art in (spoken or written) Language and Information Processing (LIP) are sound.

† I am grateful to James Allan and Fred Jelinek for inviting me to give the talk on which this paper is based, and to two referees for their comments.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

Specifically, system evaluation done within the dominant (D)ARPA paradigm is far more limited than is claimed or assumed, in relation to the realities of system use. This is not because the systems are less powerful than is desired. Rather, it follows from the nature of the *task contexts* in which systems in practice are necessarily embedded. This paper argues that we need to reconsider the dominant model of the relation between technology and its use, in which a system is simply plugged into its context of use via customisation for specific applications. This (D)ARPA-derived model is defective in itself, but is even more damaging because of the way information technology in general is evolving. Thus the paper's message is 'Don't look at the system, look at its role'. The points made in the paper may seem obvious to, for instance, the evaluation community that has been involved in the EAGLES effort (EAGLES 1996). The influence of the generic (D)ARPA model has nevertheless been large and is continuing to spread, so it is essential to recognise that while it has great value at some stages of, or for some particular purposes in, LIP research, it also has serious limitations. We should therefore look, especially for those tasks where (D)ARPA evaluation programmes have now run for some time, for new evaluation methodologies that take advantage of the (D)ARPA approach where many teams participate in a common, organised test, but which also address the critical matter of system use.

This paper therefore reviews six LIP tasks where there has been significant evaluation effort, especially under (D)ARPA auspices, namely information retrieval, speech recognition, machine translation, information extraction, automatic summarising, and intelligent inquiry. The review will examine the evaluation paradigm followed in each, the characteristic properties of the task, and the paradigm's justification and assumptions, with the aim of identifying deficiencies in the evaluation models used. I shall then consider the overall picture of (D)ARPA and (D)ARPA-style LIP evaluation that emerges, and assess this in the light of current trends in information technology, concluding with suggestions for new directions in LIP evaluation.

As the literature on these evaluations is very large, my references have been selected as useful entry points, or key citations, for the various tasks. I shall not attempt, either, to compare the (D)ARPA evaluation concepts and scenarios with those applied elsewhere, e.g. by EAGLES, or reviewed in Sparck Jones and Galliers (1996). This paper concentrates on comparisons between the different (D)ARPA evaluations and their implications.

Note on terminology

I will make use of some distinctions and terms from Sparck Jones and Galliers (1996) and EAGLES (1996). Thus, I refer not only to the components of systems and systems as wholes, but also to *setups*, i.e. systems along with their contexts, including data and people. A system performs a task which is defined by its internal *objectives*, but also serves a *function* relating to its external context. In evaluation, it is important to distinguish broad performance criteria from the specific measures that interpret them in particular ways. Evaluation properly refers both to contextual *environment variables* and to *system parameters*: these should both be visible in glass-

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

box evaluation, but parametrisation is hidden in black-box evaluation. I shall adopt the EAGLES term *usage* to refer to how systems are used by humans in setups.

2 Information Retrieval (IR)

I begin with the document/text retrieval task, where the research community's approach to evaluation is long established. The task seems simple, and the (D)ARPA/NIST Text REtrieval Conferences (TRECs), a very large scale evaluation programme within the paradigm, has now been running for nearly a decade (see Voorhees and Harman 2000; Sparck Jones 1999).

In IR, evaluation is conventionally by system effectiveness in delivering relevant documents, with respect to given requests, documents and relevance assessments, as measured by Recall and Precision. System performance in TREC is formally treated as black box, though participants' reports typically describe glass-box experiments. It seems obvious that, whatever else it may do, an IR system should be able to find relevant documents and avoid non-relevant ones, and to do this without stressing the user too much. Focusing on this basic performance requirement is, however, simplifying and abstracting by ignoring real, important detail. Some issues have to be addressed even in this reductive approach, e.g. degrees of relevance to be allowed, output comparability for different strategies, the precise form of the performance measure. But these are minor contacts with the real world compared with the other elements of the retrieval situation that are excluded, e.g. overlap in document content and document novelty, efficiency/effectiveness trade-offs, or aspects of user needs like those for particular types of document. More generally in these evaluations, simplicity in testing for the given definition of performance is achieved through controlled laboratory experiments from which the user is largely eliminated. When performance is measured, following TREC convention, by computing Average Precision over 1,000 ranks, for instance, how does this relate to the user's view of the retrieval task?

It is true that the user (or user surrogate like an information professional) is needed to make the relevance assessments, i.e. to provide the *answer data*, on which evaluation depends. However, this can be, and normally is, done in a deliberately unrealistic and even aseptic way, without any reference to retrieval as an interactive human activity. Thus assessment is done offline on the combined output of the systems tested, rather than online against individual system output, in ordinary information searching mode.

This divorce from messy reality is further illustrated by the example of figure 1. We suppose here that the user's need is for documents on 'modern laser printing technology', and we have output from two systems, A and B, that use different term weighting functions. Then we imagine that System A gives us a document on laser technology development and applications, which includes a paragraph on printing, at rank 5, and another on modern printing technology, including a paragraph on lasers, at rank 10; and we also assume that the paragraphs (and hence their host documents) are both relevant, and are also very similar. What will the user's information state and need be by rank 10, having inspected the rank 5 document?

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

```
U : 'modern laser printing technology'

SYSTEM A:  output documents

 at rank 5
   Doc 01793: 'laser technology development and applications'
     - including paragraph on printing
 at rank 10
   Doc 01462: 'modern printing technology'
     - including paragraph on lasers

    paragraphs are relevant and similar


SYSTEM B:  output documents

 at rank 5
   Doc 01462 ...
 at rank 10
   Doc 01763 ...
```

Fig. 1. Information Retrieval example – system alternatives.

Or, if we imagine the two documents reversed, as in the System B output, what again does the user think by rank 10, given what he has learnt from the documents he has already inspected? Formally measured performance after rank 10 will be the same in the two cases, though the situations are significantly different.

The laboratory-style test dominates TREC, not only in the one-off (*ad hoc*) variant of retrieval, but in others like filtering, and even in the interactive test track where, in the interests of control, real world factors like display ergonomics are excluded. This is not to suggest that controlled experiments are of no value, or that a retrieval system's internal capabilities are of no importance, or that making the user's lot easier by shifting retrieval effort to the system, are not legitimate goals. However, in TREC users are seen only as minimal data suppliers, not as the engaged information operators they really are. Interactive system evaluation is very hard, but this is not the explanation for the current style. The justification for the focus on relevant retrieved is that this is the task *core*. Serious automation means that a system should make the core contribution to its overall task, or exhibit core competence in it. In the IR case this is clearly to ensure that well-matching documents are relevant, and as this is the central system objective it should be evaluated against it. By implication this also serves to establish the system's effectiveness in serving its functional role in the larger information management setup of which it is a part.

However, concentration on the core just stimulates an ever tighter vicious evaluation circle, as the current obsession with percentage improvements in Average Precision clearly shows. So the simple observation that without users there would be no need for a retrieval system suggests that is is now time to reconsider this view of, and exclusive emphasis on, the IR core by asking:

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

1. Can we get a solid core technology?
2. If so, how solid is it by now?
3. And how important is this core in its context?

Informally, we can ask what the relative 'sizes' of core and context are: does the core loom as large as we assume, or is usage far more important that we have hitherto supposed?

Reviewing the present state of IR, the extensive evaluations already done suggest that we have converged on a respectably solid core technology, namely the statistically-based one. At the same time, performance at realistic output cutoff levels has reached a 30–40% Precision plateau for ordinary environments, with the attainable target defined by heavy-duty manual query development as only, say, 50% (Sparck Jones 2000; Gordon and Pathak 1999). These points imply that we should shift our attention to the system context and, specifically, usage. Thus we should ask what contexts are like, and how their properties bear on the way the task core is defined and approached. Context here is more than the dessicated environment familiar from traditional testing, which consists just of the supplied initial information request and the simple relevant/not relevant assessment for each separate retrieved document with respect to that request. Examining the context thus involves asking the crucial question: what are users doing in their information operations? For instance, what activities have they in hand at the time of search, why are they searching, how do they examine system output, and so forth.

3 Speech Recognition (SR)

Speech recognition, and especially continuous speech recognition, has been a major (D)ARPA concern, with high-profile evaluations since the eighties (Young and Chase 1998). As in the TREC case, the evaluations have done much to improve and consolidate SR technology; indeed the claims for the evaluations as agents (along with more machine power) in raising performance levels, and in accelerating technology development, have been much stronger than in the TREC case.

Considering these SR evaluations in the light of the comments just made on IR, it is evident that the core competence assumption in SR is very strong. If we define recognition as transcription, getting the correct transcription is the obvious core requirement; and the classic Word Error Rate (WER) performance measure, a function of insertions, deletions and substitutions in the transcription when compared with the word string actually uttered, is equally clearly appropriate. Further, it is natural with such a narrow task focus to have black-box evaluation, and controlled laboratory testing that subjects recognisers to systematic stress under varied and increasingly challenging data. As with IR, the advantages of this approach are quite evident: there has been convergence on an effective generic technology (Hidden Markov Modelling), and a significant improvement in performance through the competitive refinement that (D)ARPA-style evaluations foster.

However, when compared with IR, transcription (beyond the temporary form supporting the speech processing itself) is only at the margin as an LIP task in its

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

own right, i.e. as dictation. For speech input systems in general it is more properly a subtask, and viewing it as a subtask throws a different light on its core status. Speech-based LIP clearly cannot be done without transcription (word or sound), but what does this imply? Only that SR is necessary but not sufficient for the whole LIP task, and it is with the latter, directly or indirectly, that the user is concerned. It is indeed widely believed that for many real situations, and in particular outside very limited domains, SR technology without language understanding cannot deliver 0% WER; thus language understanding or something approximating to this is needed for correct (perhaps corrected) transcription, even if the core is limited to recognition. However, more importantly, for speech-based LIP tasks in general there is more to the core than recognition, and the language-processing element is more central to the task than the purely speech-processing one.

The current form of SR evaluation depends on a hidden assumption about task contexts, i.e. about the relation between the SR subtask component and the rest of the system, which is more than just an architectural design matter. The underlying assumption is that better subtask performance delivers better task performance. There is also, in general, a further hidden assumption, namely that this is because it is necessary, for task performance, to keep transcriptions for future use. These future uses benefit from quality transcription.

However, there is little discussion of what tasks, particularly future-use tasks, actually require. For example, spoken document retrieval may need long term records, but does not need high quality transcriptions if it is also possible to listen to the original audio. It may indeed be helpful or even necessary to do this to get features of the original speech which the transcription cannot capture. But alternatively, since listening is slow, it may be be more useful to have the transcriptions to read. However, while good quality transcriptions may be easier to read, how good they need to be depends on the purpose for which the reading is done, e.g. skimming for potential interest may not be much affected by small differences in quality. Indeed, given that there is considerable redundancy in speech, it might be rational not to aim at full transcription at all, but some reduced alternative, suggesting that WER is an inappropriate measure to start with. This applies even if transcription and cleaning up are done as distinct processing steps. Thus, as this one example shows, what is required in transcription quality has to be determined by an explicit task, and setup, analysis.

Further, the task analysis in the spoken document retrieval case suggests that WER is not an appropriate performance measure, since word order and function words are normally jettisoned in indexing: a measure like Term Error Rate (Johnson *et al.* 1999) is more suitable. Again, for the task as a whole, retrieval is the dominant element, and once fair transcription quality has been achieved the retrieval mechanism has far more impact on performance than the recogniser output. Finally, just as with text retrieval, even the task core as defined by the entire system apparatus is still only a part of the larger task setup as a whole.

With other speech-based tasks there may be no user need for transcribed records at all, so the core can look quite different. For instance, with information inquiry as illustrated by the Air Travel Information System (ATIS) task (Pallett *et al.* 1994),

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

```
DIALOGUE 1 :

U1 : Do you have trains after six ?

S1 : Yes, at 7.30 and 8.30.

U2 : That's on Sundays ?

S2 : Yes.


[U1* :  Do you have drains after six ? ]


DIALOGUE 2 :

U1 : Do you have trains <> ?

S1 : Sorry, what precisely is it you want to know ?

U2 : Trains after six on Sundays.

S2 : They are at 7.30 and 8.30.
```

Fig. 2. Speech Recognition example – system alternatives.

some correctly recognised anchors are required. However, it may not be rational to pursue accurate recognition to the limit when the information sought can be inferred, or the input clarified via dialogue, particularly when even correctly recognised inputs may be incomplete or ambiguous and call for system follow-up anyway. More generally, it is evident that good robust system design has to be hospitable and tolerant of varied user behaviour, and hence treat dialogues as dynamic wholes. System evaluation has to reflect this, implying both that conventional SR performance measures have little part to play and that, within rather broad limits, recognition quality is just one among many properties of the inputs the dialogue manager has to handle

Thus taking the example in figure 2, in Dialogue 1, where we assume that the User's utterances are perfectly transcribed in U1 and U2 as given, everything is fine. However, while we would prefer to avoid mistranscribing the user's first input as U1*, if we only partially capture it as shown in U1 in the alternative Dialogue 2, then with the system follow-up shown in S1 there, after the next User input U2 the system has the same information as in Dialogue 1 and after the same total number of turns. Similarly after S2, the user has the same information to meet their requirement in both cases.

In these retrieval and inquiry tasks the user has a key role, within the task core, though this does not apply to their SR subtasks or to some tasks embedding SR. It is thus essential to check tasks specifically for their SR requirements. Even if, in

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

general, better SR is helpful, implying some justification for the conventional mode of SR evaluation, it makes sense to assess the requirements of the task as a whole, taking usage considerations into account, before selecting SR for exclusive attention. The current state of the art is that there is some sound subtask technology, though performance is naturally condition-dependent. It is certainly good enough for some tasks but not for others, especially non-interactive ones, e.g. 'pure' translation. It is thus essential to ask how adequate this technology is for each task in its context, and in particular, in relation to where and how the user figures in the whole.

4 Machine Translation (MT)

Evaluation has been a long-standing concern in MT research, as in IR, and there is a large literature on the subject (e.g. see Falkedal 1991; Vasoncellos 1994; Sparck Jones and Galliers 1996). (D)ARPA's effort so far has been modest, and it has not had the leadership role in MT that it has in other areas like information extraction. However, since it has addressed MT evaluation (ARPA 1993; White *et al.* 1994), I shall include MT as a task within the scope of this paper, though I will consider it only briefly and from the (D)ARPA point of view. The issues of MT evaluation have been well-rehearsed, and the discussion of MT evaluation which follows is intended only to provide a summary context for comment on the (D)ARPA tests. In fact, the approach to evaluation adopted in these tests is not radically different to that used for context-independent MT evaluation elsewhere. But the attempt to apply it rigorously points up the problems with the (D)ARPA paradigm.[1]

MT evaluation has generally been approached in the same spirit as SR evaluation, though with much more complexity in the detail. Thus evaluation has been addressed with measures defining 'accuracy' or fidelity in capturing the input, e.g. identifying word senses correctly, and 'propriety' or felicity in presenting the output (e.g. delivering legitimate syntactic constructions). Evaluation is operationalised by semi-objective counting or grading measures. The assumption is that, as with SR, there is no need to consider any human contribution to the definition of the core task itself, i.e. there is a usage-free core, even if humans have to be used as pseudo-automatic evaluators for measurement purposes.

However, for a task as complex as translation, freedom from error, or perfect reproduction, is hard to operationalise. Translation is not word-for-word, and counting howlers does not go far as a performance measure. But trying to use the notion of literal translation leads to difficulties too: translation is complex and open, because language is complex, has many expressive options, and deals in inaccessible meaning, so literal translation is not well-defined. There is thus no objective base for evaluation and hence no clear system core capability. Assessment becomes a matter of reasonable or acceptable translation according to human judges, as in the (D)ARPA accuracy and propriety assessment. In the (D)ARPA evaluations, accuracy

[1] MT figures in new (D)ARPA LIP initiatives, but these are too recent for comment on their evaluation detail.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

was measured in two different ways: by adequacy on a segment by segment basis, and by informativeness in SAT-like comprehension tests; propriety was measured by fluency, judged per sentence. However, even though the adequacy and fluency assessment was quite tightly controlled, and involved grading, there is implicit reference to context, either through the tacit assumption that the highest grade is required for all purposes or that some other lower grade may nevertheless be reasonable or acceptable for some particular purpose(s). Thus on investigation the notion of core becomes suspect or variable, as with SR.

The role of context is more evident in the (D)ARPA comprehension evaluation, and becomes even clearer in the current state of MT, outside very specific applications where the need for context reference is already obvious. The general lack of quality leads to dilute performance criteria, like 'acceptability', which is implicitly context dependent: that is, even if we suppose subjective assessment by professional human judges is a legitimate form of context-free evaluation, it is currently impossible to guarantee translation of a sufficiently high quality to imply that it would be very likely to meet many contextual requirements. Reference to context and usage is therefore essential in evaluation.

The example in figure 3, using translation from English to Japanese (and also showing direct English equivalents for the latter), reinforces the point. Given the input sentence I, current translation systems have no reliable means of choosing among the alternative output sentences Oa – Oc. Oa and Ob are variations that hinge on the ambiguity of 'and' in the source language; but they also allow the alternative structural interpretations exhibited in the English equivalents, which show how the subtleties of sentence structure as indicators of relative emphasis are not necessarily easily preserved on the journey from source to target language. Oc is an equally legitimate output, even though it embodies a more substantial structural transformation of the original. At the same time, the relation between translation quality and functional role is clear when, for example, 'rough' translations of document titles from Japanese to an English which lacks articles are deemed acceptable for literature skimming, where the context requirements are much less stringent than they would be in translating commercial contracts intended to retain their legal force.

For some types of application we currently have good enough technology, but there is little generic MT technology. In the first case, context is already being factored in, explicitly or implicitly. The need for better generic technology also justifies more work on it alone, but even here references to potential contexts, their properties and in particular the role of human users within them, are useful. Developing generic technology should not be taken to imply it is possible to approximate a single core system for translation: translation is no more context-free than any other language-using task. It might be the case, if we imagine we could build what we might informally call a good, solid translator, that this would in fact be functionally satisfactory for a range of purposes. However, this would still have to be independently established, and could not be taken for granted. Just because translation is a complex task, it has complex implications for, and interactions with, its context.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

```
I : She borrowed the book and failed to return it.

Translation to Japanese:

Oa : Kanojo ha hon wo karite kaesanakatta.

[ She borrowed the book and didn't return (it).
  or
  She borrowed and didn't return the book. ]

Ob : Kanojo ha hon wo karitaga kaesanakatta.

[ She borrowed the book but didn't return (it).
  or
  She borrowed but didn't return the book. ]

Oc : Kanojo ha karita hon wo kaesanakatta.

[ She didn't return the book she borrowed. ]
```

Fig. 3. Machine Translation example – system alternatives.

5 Information Extraction (IE)

(D)ARPA has played a large role in IE evaluation, over a long series of tests in the Message Understanding Conferences (MUCs) (Sundheim 1993; Chinchor 1995, 1998; Gaizauskas and Wilks 1998). Again, these (D)ARPA evaluations have followed the SR model, with correct answers provided, though in a much more detailed and also complex form: the latter allows for partial answers, which do not figure in SR. The evaluations have applied numerical measures based on recall and precision, taken from IR but interpreted rather differently, to deliver performance scores analogous to WERs.

The notion of correctness in these evaluations is very strong because of the explicit and detailed character of the answer data, both for whole templates encompassing all the information to be extracted and for individual slot fillers. This applies especially to the earlier MUCs (to MUC-5), where the answer data was more elaborate, but also to the later ones. However, the surface appearance of the tests and resemblance to the SR case is very misleading. IE is not really like SR at all, and the (D)ARPA evaluations are very artificial and very strange. There is no reference to context, and specifically to the use(s) for which the extracted information is intended; and because there is no reference to the system's functional role, there is no rationale for the answer style. At the more detailed level the evaluations use specified generic information categories, and particular template, etc. decompositions for these, but since there is no definition of the intended usage constraints, so there is no justification for the choices of answer supplied. There is not even the sort of reference to context that user requests signal in the IR case. In the IE evaluations, whatever might be assumed about the task context is embodied, but without comment and so deeply as to be inaccessible, in the answer data; and the

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

```
I :
The company announced some developments. BoxCo, the packaging
subsidiary, would merge with Bags Inc. The subsidiary would enter
Bags Inc's Japanese market.


Oa : BoxCo to market in Japan.

Ob : BoxCo and Bags Inc to market in Japan.
```

Fig. 4. Information Extraction example – system alternatives.

answer data itself supports a variety of only weak inferences about supposed context and usage. This applies even to the later MUCs, MUC-6 and MUC-7, though the simpler forms of information to be extracted (Chinchor 1995, 1998) might imply the potential uses were more obvious.

The lack of context characterisation is, however, much less justifiable than in the SR case, where there are natural defaults, or even MT, where there is at least a weak notion of basic translation. In IE there are no natural defaults of the form: given a text, extract the essential, or all of, the facts, even when this is made subject to category guidance (e.g. terrorism) and a slot label (e.g. reason for incident). For instance, with "Ruritanian Nativists kidnapped the Grandian ambassador, citing Grandian delinquencies", was the reason that the ambassador was a Grandian, or that he was a conspicuous Grandian, or that he was a formally representative Grandian? With the (D)ARPA IE evaluations there is a presumption that the output will satisfy the intended use, but there is no independent way of verifying this. There is not even the guidance that the output will be for, say, alerting or information dissemination purposes, or for incorporation in a record file, or for entry in a conventional database, though these different functions have quite different implications for evaluation. Thus the requirements for accurate entity identification may be much less stringent in the first case than in the third, where constraints on entity integrity may apply.

Thus as an illustration, consider the input text of figure 4. Here extraction focused on information about companies in markets could emerge with either output fact Oa or output fact Ob. In an alerting service for human analysts it would probably not matter which was used, but for a formal database the choice of entity identity could have significant repercussions.

The current state of the art in IE is that there are beginning to be specialised applications, and evaluation has encouraged the development of methodologies like shallow parsing and tools like proper name recognisers. But progress with core systems has been poor, overall performance in tests is only moderate, of order 50%, and there is a lack of coverage in difficult areas, e.g. the extraction of temporal relations. However the issue is not primarily system deficiencies that can be overcome: the problem is that the notion of core, independent of applications, is weak. Performance has been best on (sub)tasks like named entity recognition, not just because they are easier but also because they are relatively uncontroversial. Performance is more limited for subtler subtasks, such as co-reference identification

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

and predicate-argument structure extraction, not only because they are operationally more difficult but because of disagreement about the answer information, i.e. about the precise nature of the type of information to be extracted (and hence about its instances). However, this situation is not surprising because the notion of common system core is weak: IE is intrinsically context dependent. It requires a functional specification, particularly in relation to user requirements.

I shall consider the remaining two tasks more briefly, as there has been less systematic evaluation for them, both generally and within the (D)ARPA paradigm.

6 Automatic Summarising (AS)

It might appear that there is as much of a core in AS as there is in MT, because there is an analogy with literal translation in the notion of a straightforward reduction of a text to its essentials. This supplies a basis for core evaluation.

However, even without the problems that evaluating literal translation involves, the analogy does not hold because summarising is a fundamentally much more challenging task, involving a qualitative difference in the input/output transformation relation. In AS, extended discourse has to be condensed, implying that the source text has to be considered as a whole, where MT, in contrast, is locally bounded. So trying to evaluate by analogy with the MT case is not productive. But for the task's primary reduction operation, an evaluation criterion of the form: Have the important input concepts been captured? is far too loose; and the same applies to a criterion for its secondary production operation like: Is the output text coherent? Even trying to evaluate AS in a more limited way by asking whether a system has extracted some specific concept, independently deemed important, from the source is problematic because there are so many expressive possibilities for concepts. Information condensation is not information extraction, and because condensation is required, AS is much more complex and a much more open task even than IE.

The only form of core-confined evaluation that would appear to be available for AS is thus the very weak one of acceptability, according to some human judges armed with what will necessarily be rather broad guidelines. However, this is not in fact context-free. It involves hidden assumptions, and stiffer ones than are needed in the analogous MT case. Thus, without more precise direction, acceptability evaluation for AS takes reflective summarising as the default, i.e. assumes that the destination community for the summary has the same properties as that for the source text with respect to interests, technical background, language 'bias', and so forth. The idea of a default summary as reflective of source is in fact importing functional considerations into the evaluation, without checking their validity or applying them directly in the evaluation. There is also a tacit presumption that the purpose for which summaries are used is for guidance on the source text.

However, even if reflective summarising is a legitimate goal, evaluation by acceptability for AS is extremely unconstrained, much more so than in the analogous MT case. A summary is a radical transformation of its source, implying far more possible output alternatives than in the relatively more limited MT situation. There is nothing like a natural summary for a text, even a natural reflective summary,

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

```
I :
The tests were conducted on a thousand wombats. They showed half
preferred shredded wheat, half rice crispies.


Oa : Wombats have no cereal preferences.

Ob : Wombats will eat breakfast cereals.

Oc : Wombats are like people.

Od : Huge test on wombats.
```

Fig. 5. Automatic Summarising example – system alternatives.

so there cannot be any effective, autonomous core evaluation. It is necessary to consider the context and usage for summaries. This makes AS very unlike SR, and more like IR.

These realities of AS are illustrated in figure 5. Given the example input text, I, shown there, there are no independent reasons for regarding any one of the output summaries Oa - Od as superior to the others, though they are quite different; and it is hard to exclude any as less legitimate through relying on world knowledge not directly conveyed by the text, since there is no requirement – or indeed, possibility in general – of delivering quality summaries without this.

Research on automatic summarising, as opposed to text extraction, has barely begun, and because non-extractive summarising is very hard, work on core technology that would be useful for many purposes is valuable. However, both principle and practice imply that it is essential for AS to attend to its context. Indeed, for this task (D)ARPA itself has had to concede more context reference than is normal outside IR, where there can be no evaluation without reference to users. Thus in the (D)ARPA evaluation (Firmin and Chrzanowski 1999; Mani *et al.* 1999), summaries were evaluated for their capacity to support external tasks, i.e. extrinsically, as well as intrinsically. Specifically, intrinsic summary acceptability was complemented by extrinsic summary effectiveness for various purposes: in helping the user in information retrieval by offering a convenient base for assessing likely document relevance, in facilitating document categorisation, and also for question answering. However the first two tasks imposed quite weak demands on summarisers, especially in relation to the textual properties of system outputs and, more importantly, details of their environments were not fully and explicitly registered.

7 Intelligent Inquiry (II)

I shall consider this concluding task only very briefly too since evaluation, particularly within the (D)ARPA programme, has been quite limited. Further, intelligent inquiry is really a label for a family of tasks, taken together here only because they involve extended dialogue.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

```
DIALOGUE 1 :

U1 : Do you have trains after six in the evening ?

S1 : Yes, at 7.30pm and 8.30pm.

U1 : On Sundays ?

S2 : Yes

U3 : Before six ?


S3a : At 8.00am.

S3b : Only very early, at 8.00am.
```

Fig. 6. Intelligent Inquiry example – system alternatives.

I considered intelligent inquiry under SR in relation to the 'scope' of performance computation, and also mentioned there the ATIS domain used in the (D)ARPA evaluations. I am returning to these evaluations here because, as with IR, the user is essential, i.e. the user is within the core. Thus assessing the performance of an inquiry system on a paired question-answer (QA) basis, and characterising satisfactory answers on the IE model in relation to their immediately prompting questions, as in the official (D)ARPA tests (Pallett *et al.* 1993, 1994), will only go so far in evaluation. The core capability demonstrated is relatively limited because of the constraints imposed by the close relationship between question and answer. In reality, interaction in dialogue is discourse dependent in a fuller sense, and communicative behaviour is adaptive. An II system needs to support constructive, i.e. contextually-motivated dialogue, implying system development that attends to context and focuses on the user in it. This also implies an expensive form of evaluation, with many users and, as in interactive IR, session-level performance evaluation.

But the need for this approach is illustrated by figure 6, using a variation on the SR example dialogue used earlier. Thus, if we have the sequence up to S2 followed by the user input U3, while S3a is satisfactory on a QA basis, responding to the user's change of immediate goal, S3b is a response that recognises the user's dominant goal of travelling later rather than earlier in the day. However, in the present state of the art it is still reasonable to work on the first-level QA technology, as long as its restrictions in relation to context considerations are recognised.

8 Comparative conclusion on the tasks

Drawing together the comments on the individual tasks and comparing them, it is evident that the shared predominant (D)ARPA style, without or with minimised

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

context reference, is very misleading. Beneath the surface similarity reflected in the recurrence of 'recall' and 'precision' across the programmes, there are significant differences between the tasks. These are:

- in the nature of the task core, and hence what core system or core technology mean;
- in the status of the current core technology in relation to what is desirable or attainable;
- in the contribution of the user even to basic core evaluation;
- in the balance between core and context, i.e. what core competence contributes to the overall task.

Thus, for example,

- MT is nearer to having an autonomous core than IE;
- IR current core is much better than AS current core;
- IR core evaluation requires the user in the way MT does not;
- in AS and MT, system core competence contributes more to overall task attainment than in IR (other than in vulgar file manipulation).

The overall message from the review is that you must consider the context and user not just when you only have *some* core technology and need assistance to compensate for its deficiencies, but equally when you have *established* core technology. It is necessary to examine context, especially the user, not simply because this is helpful, but because LIP tasks always demand it, later if not sooner. The context matters regardless of the 'size' of the core: a *correct core* is in principle as well as practice unattainable.

9 The future, and the impact of IT evolution

Turning now to evaluation in the future, there is of course plenty of scope for technology evaluation: we need core capabilities and should work on technologies for them, where the current style of evaluation is useful. However, it is essential to be clear about what the task core is, and what its limits must be. Even here, moreover, it is desirable to bring context in sooner than is nowadays usual, at least through careful, detailed, more far-reaching environment specifications, i.e. through clear characterisations of environment variables and their values, with systematic testing against these under serious experimental control. Bringing context in fully then implies involving real users; and for those tasks where we already have some non-trivial technology, this should be done now.

This user involvement is not just desirable in its own right. It is becoming increasingly important given the impact of IT evolution, notably that of the Web. In this new situation, LIP tasks are not disjoint, to be evaluated in stand-alone mode, offline. Information management is seamless. Thus considering the information actions that are natural in these new IT environments, with their many different tools and resources, it is evident that users mix information actions opportunistically, treating particular tasks as subtasks for other tasks, and moving freely between

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

(sub)tasks. For instance in the IR case, conventional formal performance measures (and hence search engine quality) become irrelevant if the user moves quickly away from the initial search output on to Web pages and pursues their links. Tasks are left incomplete as users redefine their goals. This implies e.g. highlighting 'summary' sentences on the fly for unique situations, and similarly IE or MT tailored to the current context (as already, in a simple way, in cross-language IR). In the new IT world, dynamic context is crucial and it needs tackling in evaluation.

Unfortunately, this has the consequence that (sub)task specification becomes harder: for instance what does such 'multi-purpose' usage demand of the IE core, what should the extracted fact database intended for varied use (and misuse) be like? Equally, user-oriented evaluation becomes much harder. It is a challenge even to capture the system environment in a relatively limited sense, and certainly to evaluate whole setups. Thus, what is an effective *information management* system or setup? In the new world there are, in particular, two jokers: first, generalising over applications, i.e. defining generic tasks; and second, concatenating less than perfect systems without intervening users, as may well occur. If the output from one task is the input to another, as an intermediate step in response to a user demand, there could be incremental degradation. At least it is not clear that one component will automatically compensate for the deficiencies of another; or that it is at all easy to build components that can compensate at once for different types of deficiency in inputs from different sources, as could be required of a information extraction module taking its input from any of a recogniser, a translator, or a summariser. We have a new problem of the indirect, rather than direct, user. This then leads to a new form of the original evaluation need, namely how to evaluate subtasks just as components, but in this case necessarily without the leverage of usage, rather than just without bothering about it.

Overall, IT evolution pulls LIP system evaluation in opposite directions. Better systems need stronger user involvement. But since these will also usually be more complex systems, this implies not only a need for development evaluation without too much hassle and cost: it also implies a need to evaluate subsystems which may have no direct user relation or meaning. For example, there is no reason to suppose that the ordinary information system user would interact with the output of a syntactic analysis phase in text interpretation. So this has to be evaluated in some necessarily artificial user-independent way. However, even though the two forms of evaluation are both wanted, and are not readily reconciled, my argument is that the (D)ARPA programmes have overemphasised the second form of evaluation, and we now need to devote more attention to the first, however hard it is.

References

ARPA (1993) *Report of the Advanced Research Projects Agency Machine Translation Program, System Evaluation*. `http://ursula.georgetown.edu/`.

Chinchor, N. (1995) Overview of results of the MUC-6 evaluation. *Sixth Message Understanding Conference (MUC-6)*, pp. 13–31. San Francisco: Morgan Kaufmann.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

Chinchor, N. (1998) Overview of MUC-7/MET-2. =*Message Understanding Conference Proceedings: MUC-7*. http://www.muc.saic.com/proceedings/muc-7-toc.html.

EAGLES (1996) *EAGLES: Evaluation of Natural Language Processing Systems*. Final Report, EAGLES Document EAG-EWG-PR2. (From Centre for Sprogteknologi, Njalsgade 80, 2300 Copenhagen, Denmark.)

Falkedal, K. (1991) *Evaluation Methods for Machine Translation Systems: An Historical Overview and Critical Account*. Report, ISSCO, Université de Genève.

Firmin, T. and Chrzanowski, M. J. (1999) An evaluation of automatic text summarisation systems. In: I. Mani and M. T. Maybury (eds.), *Advances in Automatic Text Summarisation*, pp. 325–339. Cambridge, MA: MIT Press.

Gaizauskas, R. and Wilks, Y. (1998) Information extraction: beyond document retrieval. *J. Documentation*, **54**(1), 70–105.

Gordon, M. and Pathak, P. (1999) Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Infor. Process. and Manage.*, **35**(2), 141–180.

Grishman, R. and Sundheim, B. (1995) Design of the MUC-6 evaluation. *Sixth Message Understanding Conference (MUC-6)*, 1–11. San Francisco: Morgan Kaufmann.

Johnson, S. E., Jourlin, P., Moore, G. L., Sparck Jones, K. and Woodland, P. C. (1999) Spoken document retrieval for TREC-7 at Cambridge University. In: E. M. Voorhees and D. K. Harman (eds.), *Seventh Text REtrieval Conference (TREC-7)*, 191–200. (Special Publication 500-242, Gaithersburg, MD: National Institute of Standards and Technology.)

Mani, I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T., Chrzanowski, M. and Sundheim, B. (1998) *The TIPSTER SUMMAC Text Summarisation Evaluation*. Final Report, The Mitre Corporation, Washington C3 Center, McClean, VA.

MUC-5 (1993) *Fifth Message Understanding Conference (MUC-5)*. San Francisco: Morgan Kaufmann.

MUC-6 (1995) *Sixth Message Understanding Conference (MUC-6)*. San Francisco: Morgan Kaufmann.

MUC-7 (1998) *Message Understanding Conference Proceedings: MUC-7*. http://www.muc.saic.com/proceedings/muc-7-toc.html.

Pallett, D. S., Fiscus, J. G., Fisher, W. M. and Garofolo, J. S. (1993) Benchmark tests for the DARPA spoken language programme. *Human Language Technology: Proceedings of a Workshop*, 7–18. San Francisco: Morgan Kaufmann.

Pallett, D. S., Fiscus, J. G., Fisher, W. M., Garofolo, J. S., Lund, B. A. and Przybocki, M. A. (1994) 1993 benchmark tests for the ARPA spoken language programme. *Proceedings Human Language Technology Workshop*, 49–74. San Francisco: Morgan Kaufmann.

Sparck Jones, K. (1999) Further reflections on TREC. *Infor. Process. and Manage.*, **36**(1), 37–85.

Sparck Jones, K. (2000) Summary performance comparisons: TREC-2 through TREC-8. In: Voorhees, E. M. and Harman, D. K. (eds.), *Eighth Text REtrieval Conference (TREC-8)*.

Sparck Jones, K. and Galliers, J. (1996) *Evaluating Natural Language Processing Systems: Lecture Notes in Artificial Intelligence 1083*. Springer-Verlag.

Sundheim, B. (1993) Tipster/MUC-5 information extraction evaluation. *Fifth Message Understanding Conference (MUC-5)*, 27–44. San Francisco: Morgan Kaufmann.

Vasoncellos. M. (ed.) (1994) *MT Evaluation: Basis for Future Directions: Proceedings Workshop Sponsored by the National Science Foundation*. Washington, DC: Association for Machine Translation in the Americas.

Voorhees, E. M. and Harman, D. K. (1999) Overview of the Seventh Text REtrieval Conference (TREC-7). In: Voorhees, E. M. and Harman, D. K. (eds.) *Seventh Text REtrieval Conference (TREC-7)*. pp. 1–23. Special Publication 500-242, Gaithersburg, MD: National Institute of Standards and Technology.

Voorhees, E. M. and Harman, D. K. (2000) Overview of the Eighth Text REtrieval Conference (TREC-8). In: Voorhees, E. M. and Harman, D. K. (eds.) *8th Text REtrieval Conference (TREC-8)*.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press

White, J. S., O'Connell, T. and O'Mara, F. E. (1994) *Advanced Research Projects Agency Machine Translation Program: 3Q94 Evaluation.* `http://ursula.georgetown.edu/`.
Young, S. J. and Chase, L. L. (1998) Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes. *Computer Speech & Lang.*, **12**, 263–279.

https://doi.org/10.1017/S1351324901002583 Published online by Cambridge University Press