Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Artificial Intelligence 144 (2003) 41–58

Artificial Intelligence

www.elsevier.com/locate/artint

Lessons from a failure: Generating tailored smoking cessation letters

Ehud Reiter [a,*], Roma Robertson [a,b], Liesl M. Osman [c]

[a] *Department of Computing Science, University of Aberdeen, Aberdeen, UK*
[b] *Department of General Practice and Primary Care, University of Aberdeen, Aberdeen, UK*
[c] *Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, UK*

Received 15 August 2001; received in revised form 12 July 2002

Abstract

STOP is a Natural Language Generation (NLG) system that generates short tailored smoking cessation letters, based on responses to a four-page smoking questionnaire. A clinical trial with 2553 smokers showed that STOP was not effective; that is, recipients of a non-tailored letter were as likely to stop smoking as recipients of a tailored letter. In this paper we describe the STOP system and clinical trial. Although it is rare for AI papers to present negative results, we believe that useful lessons can be learned from STOP. We also believe that the AI community as a whole could benefit from considering the issue of how, when, and why negative results should be reported; certainly a major difference between AI and more established fields such as medicine is that very few AI papers report negative results.

Keywords: Natural language processing; Natural language generation; Knowledge acquisition; User modelling; AI and Medicine; Smoking cessation; Evaluation; AI methodology; Clinical trials; Negative results

1. Introduction

The AI community is reluctant to talk about failures. While papers describing refuted hypotheses are common in more mature fields such as physics and medicine, they are rare in artificial intelligence journals and conferences, including the *Artificial Intelligence* journal. This is a pity, because negative results are important for theory formation; as with

* Corresponding author.
E-mail addresses: ereiter@csd.abdn.ac.uk (E. Reiter), roma.robertson@abdn.ac.uk (R. Robertson), l.osman@abdn.ac.uk (L.M. Osman).

0004-3702/02/$ – see front matter
doi:10.1016/S0004-3702(02)00370-3

any other kind of learning exercise, theory induction works better if it has access to negative as well as positive examples. As Drew McDermott wrote many years ago [20], 'AI as a field is starving for a few carefully documented failures'.

In this spirit we in this paper discuss the STOP project. STOP was a natural language generation (NLG) system that generated tailored smoking cessation letters. Tailoring was based on a 4-page multiple-choice questionnaire about the smoker's habits, health, concerns, and so forth. STOP was evaluated in a randomised controlled clinical trial; this showed that STOP's tailored letters were no more effective than control non-tailored letters.

The rest of this paper is divided into two parts. In Sections 2–4, we present STOP as a case study, describing the application, the KA and NLG techniques used to build the system, and the clinical evaluation of the system. In Section 5, we discuss more generally the methodological issues of what can be learned from negative results, and why so few such results are published in AI.

This article is intended for computer science audiences. Medical audiences interested in STOP should read our paper in the *British Medical Journal* [18].

2. Background and previous work

2.1. The application: Computer-tailored patient information

The most important influence on many people's health is their behaviour, including smoking, diet, and compliance with treatment regimes. This has led to increased interest in IT systems which help people change their behaviour in medically useful ways, such as stopping smoking, by providing them with appropriate information. Of course it is very difficult to give up smoking, and simply providing information is unlikely to help most people stop smoking. But clinical trials have shown that brief (5 minutes or less) one-off discussions about smoking with a doctor can help approximately 2% of smokers to quit [17]. While a cessation rate of 2% is not high, it is still very useful from a medical perspective if it can be achieved cheaply. Unfortunately, few doctors routinely have such discussions with patients, in part because of time pressure, and also perhaps in part because it is difficult for a human to remain enthusiastic about a technique with a 98% failure rate. Since IT systems do not get discouraged by high failure rates, an IT system which helped a similar percentage of smokers to quit would be very useful to the medical community.

Over the past ten years a number of computer systems have been developed that produce tailored texts (typically letters, leaflets, or manuals) for patients, with the goal of helping at least a small number of them change their behaviour. Tailoring is typically based on a paper questionnaire or telephone interview. Several of these systems are described in a special issue (volume 36, number 2) of *Patient Education and Counselling*, which also includes a review article by Strecher [35] that summarises projects in the smoking cessation area. Although Strecher concludes his review article on a positive note, in fact the experiments he reviews do not unambiguously demonstrate the effectiveness of short stand-alone tailored texts for smoking cessation (see discussion in [18]). For example, the study which is most similar to STOP, Experiment 1 in Strecher et al. [36], used a very small number of subjects (72 vs. 2553 in STOP), which meant that its statistical significance claims are

highly dependent on the exact statistical analysis and demographic adjustment techniques used.

Most previous work in generating tailored patient texts has been carried out in the medical community, and has been based on simple string-concatenation ('mail-merge') technology. Within the AI/Medicine community, a few projects have investigated using AI techniques for generating tailored patient information ([4] reviews much of this work), although we are not aware of any such projects in the smoking cessation area. For example, PIGLIT [5] generated tailored hypertext information documents that communicated information about diseases and treatments, where the tailoring was based on a patient's medical record. A clinical evaluation of one version of PIGLIT measured PIGLIT's effect on patient satisfaction and patient anxiety; it showed a statistically significant effect on satisfaction, but not on anxiety [8,15].

2.2. Evaluation of NLG systems

Perhaps the most unusual aspect of STOP from an NLG perspective was that the system was evaluated in a randomised controlled clinical trial. To the best of our knowledge, no other NLG system has been so evaluated. The PIGLIT evaluation mentioned above comes closest, but in this evaluation some of the tailored texts were manually authored by the experimenters, who simulated the NLG algorithm; all of the tailored texts in the STOP clinical trial were produced by the STOP software.

Mellish and Dale's review article of NLG evaluation [21] argues that evaluations of NLG systems can focus on the effectiveness of underlying theories, on general properties of NLG systems and texts (such as computational speed, or text understandability), or on the effectiveness of the generated texts in an actual task or application context. A clinical trial is a type of task-effectiveness evaluation, which is the most expensive and difficult-to-organise type of evaluation. In fact, until recently such evaluations of NLG systems have been rare and have sometimes suffered from major methodological problems (for example, the IDAS evaluation [19] did not include a control group). However, over the past few years a number of groups have conducted and reported reasonably rigorous task evaluations of NLG systems (for example, [7,10,13,38]). This growing interest in task evaluations may reflect the fact that such evaluations are the only known way to measure the effectiveness of NLG systems with real users. If the results of cheaper and simpler evaluation techniques such as corpus-based evaluation [3] were known to be correlated with the results of task-based evaluations in NLG (as has been suggested for machine translation [23]), then perhaps these techniques could be used to evaluate system effectiveness. However, no evidence of such correlation has yet been presented for the evaluation of NLG systems.

Task-effectiveness evaluations with real users have perhaps been unusual in other areas of AI as well. For example, in August 1996 the *Artificial Intelligence* journal published a special issue on Empirical Methods in AI. While all eleven papers in this special issue included some type of evaluation, only one paper [33] actually evaluated the task performance of users; the other ten papers used either computational or corpus-based evaluations.

3. The STOP system

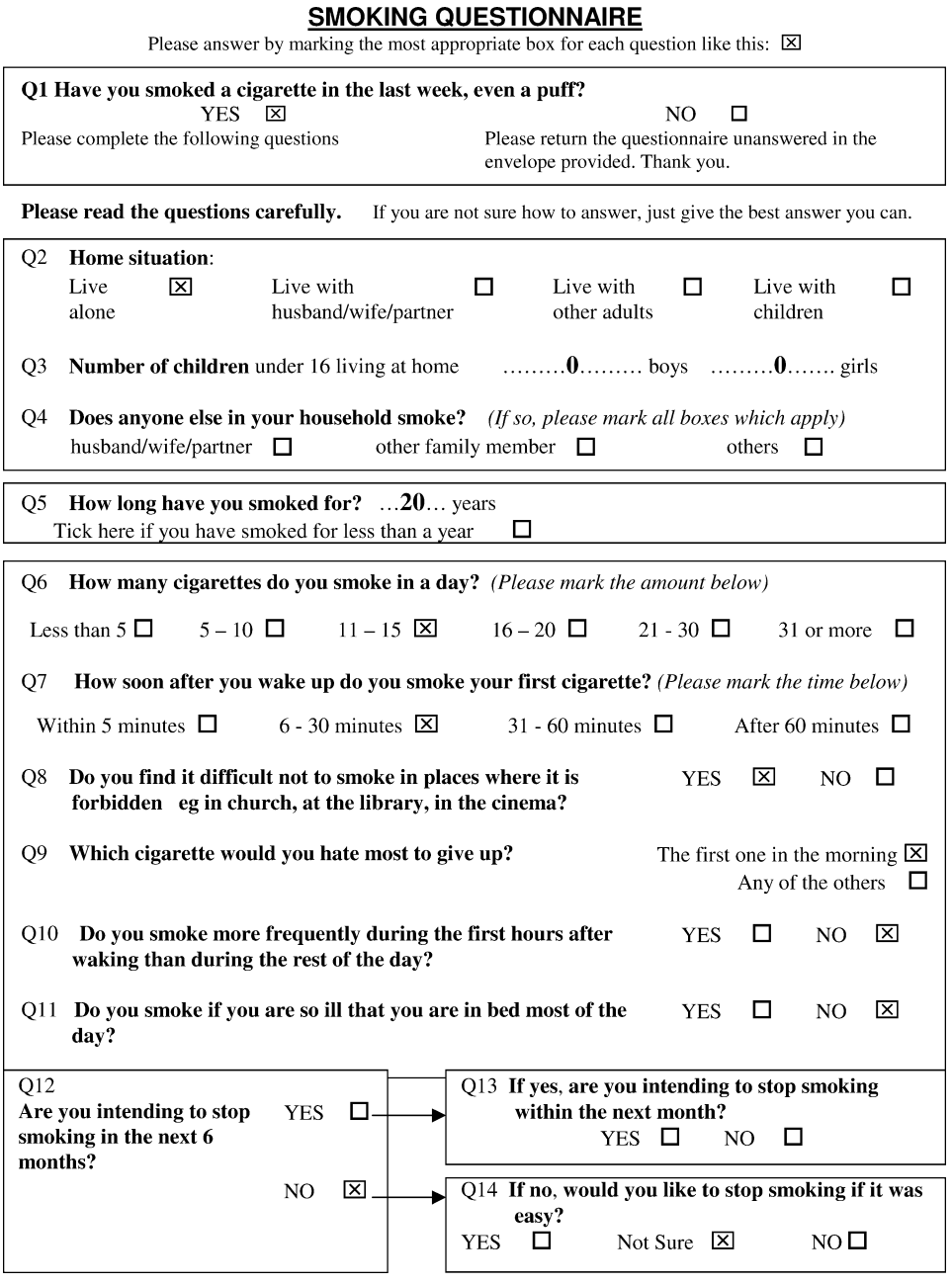
The STOP system produced a small (4 pages of A5) smoking cessation letter, based on a smoker's responses in a 4-page questionnaire. The first page of a questionnaire from a real smoker, Heather Stewart, is shown in Fig. 1. In order to preserve patient confidentiality, we have changed the smoker's name, only shown part of her questionnaire, and typed the smoker's hand-written responses. STOP also gets some basic information, such as age and sex, from the patient's medical record. Fig. 2 shows the two inside pages of the letter generated by STOP for Heather Stewart (the front and back pages are not heavily tailored, so we have not shown them).

3.1. Implementation

STOP questionnaires were read by an optical scanner, and stored in a Microsoft Access database. STOP could also be accessed via a Web front-end, in which case questionnaires were entered via an HTML form and transmitted to STOP via CGI. In either case, the questionnaire was processed by the core STOP system, described below. This system produced either an RTF file, which was printed with Microsoft Word, or an HTML file, which was displayed using a Web browser.

STOP produced a letter of four A5 pages. The front page of a letter contained the names of the smoker and of the smoker's GP (General Practitioner) medical practice, and an introductory paragraph. The back page was selected from one of sixteen possible back pages, but was not tailored in detail. The two inside pages were fully tailored. The tailored parts of STOP letters (the introductory paragraph and the inside pages) were produced by a fairly standard NLG system which followed the three-stage pipeline model [29]. Processing was divided into the stages of document planning, microplanning, and realisation, of which document planning (deciding what information to communicate) was the most complex. Essentially, the document planner worked by classifying smokers into one of 7 categories, and then running a high-level category-specific schema that specified which sections and paragraphs should be included in the letter. The category schema also specified the importance of different sections and paragraphs, which influenced their length. Detailed decisions about the content of individual paragraphs were made by smaller schemas, and were again based on questionnaire information. This combination of classification and schemas is similar in concept to the Exemplars system [37], although the implementation was different. Like Exemplars, STOP was implemented in Java.

The document-planning schemas produced a tree, known as a document plan. Each leaf node of the tree defined one sentence or phrase in the letter. The internal nodes of the tree indicated how sentences and phrases were grouped, associated document structures (such as paragraphs or itemised lists) with such groups, and sometimes specified connectives (cue words for discourse relations, such as *However*) between daughter nodes. Sentences and phrases were represented by what Reiter and Dale [29] call 'canned text'; that is, strings which fully specify word forms and word order, but do not fully specify capitalisation, punctuation, and inter-token white space. The STOP microplanner and realiser converted this structure to a Word RTF document or an HTML Web page. Further details and



SMOKING QUESTIONNAIRE

Please answer by marking the most appropriate box for each question like this: ☒

Q1 Have you smoked a cigarette in the last week, even a puff?

YES ☒ NO ☐

Please complete the following questions

Please return the questionnaire unanswered in the envelope provided. Thank you.

Please read the questions carefully. If you are not sure how to answer, just give the best answer you can.

Q2 **Home situation**:

Live alone ☒ Live with husband/wife/partner ☐ Live with other adults ☐ Live with children ☐

Q3 **Number of children** under 16 living at home ……**0**…… boys ……**0**……. girls

Q4 **Does anyone else in your household smoke?** *(If so, please mark all boxes which apply)*

husband/wife/partner ☐ other family member ☐ others ☐

Q5 **How long have you smoked for?** …**20**… years

Tick here if you have smoked for less than a year ☐

Q6 **How many cigarettes do you smoke in a day?** *(Please mark the amount below)*

Less than 5 ☐ 5 – 10 ☐ 11 – 15 ☒ 16 – 20 ☐ 21 - 30 ☐ 31 or more ☐

Q7 **How soon after you wake up do you smoke your first cigarette?** *(Please mark the time below)*

Within 5 minutes ☐ 6 - 30 minutes ☒ 31 - 60 minutes ☐ After 60 minutes ☐

Q8 **Do you find it difficult not to smoke in places where it is forbidden eg in church, at the library, in the cinema?** YES ☒ NO ☐

Q9 **Which cigarette would you hate most to give up?** The first one in the morning ☒ Any of the others ☐

Q10 **Do you smoke more frequently during the first hours after waking than during the rest of the day?** YES ☐ NO ☒

Q11 **Do you smoke if you are so ill that you are in bed most of the day?** YES ☐ NO ☒

Q12 **Are you intending to stop smoking in the next 6 months?** YES ☐ NO ☒

Q13 **If yes, are you intending to stop smoking within the next month?** YES ☐ NO ☐

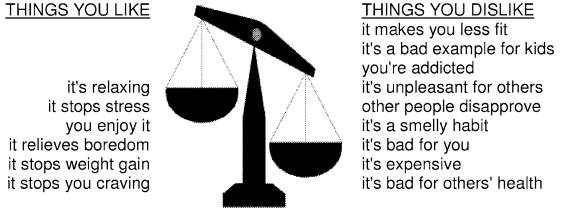
Q14 **If no, would you like to stop smoking if it was easy?** YES ☐ Not Sure ☒ NO ☐

Fig. 1. First page of Heather Stewart's questionnaire.

Smoking Information for Heather Stewart

You have good reasons to stop...

People stop smoking when they really want to stop. It is encouraging that you have many good reasons for stopping. The scales show the good and bad things about smoking for you. They are tipped in your favour.



You could do it...

Most people who really want to stop eventually succeed. In fact, 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected.

Although you don't feel confident that you would be able to stop if you were to try, you have several things in your favour.

- You have stopped before for more than a month.
- You have good reasons for stopping smoking.
- You expect support from your family, your friends, and your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.

Overcoming your barriers to stopping...

You said in your questionnaire that you might find it difficult to stop because smoking helps you cope with *stress*. Many people think that cigarettes help them cope with stress. However, taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking. There are some ideas about coping with stress on the back page of this leaflet.

You also said that you might find it difficult to stop because you would *put on weight*. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise.

And finally...

We hope this letter will help you feel more confident about giving up cigarettes. If you have a go, you have a real chance of succeeding.

With best wishes,

The Health Centre.

Fig. 2. Inside pages of letter generated for Heather Stewart.

explanatory rationale about STOP's architecture and data structures are given by Reiter and Robertson [27,30].

For example, consider Heather Stewart, for whom STOP generates the letter shown in Fig. 2. Based on her questionnaire data, Heather is classified as 'Category 3', which is a smoker who appears to dislike smoking but is not currently intending to try to stop smoking. The high-level schema for this category specifies that the inside pages should include a section on decisional balance (pros and cons of smoking), on confidence-building, and on barriers to stopping, along with a short conclusion. No information about stopping techniques is given to smokers in this category. The detailed schemas flesh out these sections. For instance, the decisional balance schema notices that Heather dislikes more things about smoking than she likes, and builds this section around a graphic which emphasises this point.

Perhaps the most innovative aspect of STOP from an NLG technology perspective was its use of revision to optimise the content of a letter, given the size constraint (four pages of A5). In general terms, this was done by having schemas annotate document plan constituents with importance markers. If the letter was too long, a revision module deleted the least important constituents until the size limit was satisfied. Reiter [28] gives further details about this process.

3.2. The non-tailored letter

As part of the clinical trial, we wanted to compare cessation rates in a group which received tailored letters to cessation rates in a group which received a non-tailored letter which was as similar as possible to the tailored letters. We decided to base this non-tailored letter on the default rules in STOP. In order to make STOP robust, the development team wrote default code to handle cases where questionnaires were incomplete, inconsistent, or illegible (to the scanner); the purpose of such code was to enable the system to say something generally useful about a topic when tailoring was not possible because of questionnaire problems. The non-tailored letter was basically an edited and tidied-up version of the letter produced by STOP when all these defaults were activated, that is when the system was given a questionnaire with no data.

The two inside pages of the non-tailored letter are shown in Fig. 3. It contains a section on decisional balance, a small confidence-building section, and a section (with three subsections) giving advice on how to stop. There is no section on barriers to stopping (such as the *Overcoming your barriers to stopping* section in Fig. 2, Heather Stewart's letter), because it is difficult to write a non-tailored barriers section. The contents of sections in the non-tailored letter are generally at least broadly similar to the contents of sections in the tailored letters; for example, the decisional balance section in the non-tailored letter uses a similar graphic to the one in Heather Stewart's tailored letter.

3.3. Knowledge acquisition

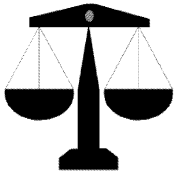
A system such as STOP must of course be based on extensive knowledge about smoking, behaviour change, and effective writing. As far as we can tell from published papers, most previous systems which produced tailored smoking cessation material have been

Information for Stopping Smoking

Do you want to stop smoking?

Everyone has things they like and dislike about their smoking. The decision to stop smoking depends on the things you don't like being more important than the things you do like. It can be useful to think of it as a balance. Have a look on the scales. What are the good and bad things for you?

GOOD THINGS	BAD THINGS
you enjoy it	it's bad for you
it's relaxing	it makes you less fit
it stops stress	it's expensive
it breaks up the day	it's a bad example for kids
it relieves boredom	it's bad for others' health
it's sociable	you're addicted
it stops weight gain	it's unpleasant for others
it stops you craving	other people disapprove
	it's a smelly habit



Add any more that you can think of. Are you ready to stop smoking? If yes, maybe it's the right time to have a go. If no, think about the good and bad things about smoking. This might swing the balance for you.

You can do it.....

People who want to stop smoking usually succeed. 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected!

Try it out.....

If you don't feel ready for an all-out attempt to stop smoking, there are some useful ways to prepare yourself. You could try some of the following ideas now. This will help you when you try to stop smoking.

- Delay your first cigarette of the day by half an hour.
- Stop smoking for 24 hours.
- Cut down the number you smoke by 5 cigarettes per day.

Planning will help.....

When you stop, it helps to plan ahead. Here are some things that have worked for others:

- Pick a day to stop, and let your family and friends know.
- Think of situations where you might feel tempted to smoke, and plan how you could avoid or deal with them.
- Get rid of all cigarettes and ashtrays the day before.
- When you do stop, take one day at a time; don't look too far ahead.

If it gets tough.....

Many people do hit rough patches; there are ways to deal with these. On the back page are some suggestions that other people have found useful.

If you do have a cigarette after a few days just put it behind you and keep on trying. Prepare yourself for another attempt, many people have more than one go before they stop for good!

With best wishes.

The Health Centre.

Fig. 3. Inside pages of non-tailored letter.

theory-driven; that is, their knowledge primarily came from psychological theories of behaviour change such as Stages of Change [26]. STOP, in contrast, made some use of such theoretical models but was primarily based on knowledge acquired from practitioners (three doctors, a nurse, and a health psychologist) using fairly standard expert-systems knowledge-acquisition techniques [6,34]. In particular, we used sorting (where our experts divided sets of smoker questionnaires into subsets) to develop our 7 high-level smoker categories; and think-aloud protocols (where our experts wrote example letters from questionnaires, and thought aloud into a tape recorder while they did so) to develop our content schemas.

A detailed description of our KA techniques, including some evaluation of their effectiveness, is given in [32]. In general terms our evaluation of KA showed that our KA techniques were good at suggesting possible categories and schemas, but that we had probably made some mistakes when we decided which of these to actually include in STOP. For example, the sorting exercises primarily suggested categorising smokers by their intention to quit and their desire to quit. However, some experts in the sorting exercises also partially categorised smokers on the basis of difficulty of quitting (for example, whether the smoker was addicted to nicotine). We decided to base STOP's smoker categories purely on intention and desire to quit; a later analysis [32] suggested that ignoring difficulty of quitting was perhaps a mistake.

Our analysis of KA effectiveness also suggested that our experts may have missed some useful tailoring rules. For example, all of the example letters written in the think-aloud protocols were fairly non-technical, and did not give detailed medical information about smoking; hence all of the letters produced by STOP were similarly non-technical. This is probably the right approach for most smokers, but a few individuals said they would have liked more in-depth medical information. Since the whole point of tailoring is to cater to individual preferences, STOP should have provided such information if the smoker desired it (this could have been a question on the questionnaire); but we did not implement this because we did not observe it in our KA sessions. Similarly STOP's letters tried to emphasise the positive benefits of quitting and avoided negative scare tactics, because this is what our experts did; and again while this is probably right for most smokers, a few individuals commented that a more 'brutal' approach would work better for them.

Some of these problems could have been due to the fact that the experts we consulted were not in fact experts on writing tailored smoking cessation letters. This is because there are no experts at this task, since manually writing tailored smoking cessation letters is too expensive to be practical. The doctors and the nurse were experts on oral consultations with smokers, and the health psychologist was an expert on writing non-tailored health information letters, but none of them had previously written tailored smoking cessation letters.

4. The clinical trial

The STOP clinical trial is described in detail by Lennox et al. [18]; here we give a brief summary for computing science (as opposed to medical) audiences.

4.1. Organisation

The trial was organised as follows. We contacted 7427 smokers, and asked them to participate in the trial. 2553 smokers agreed to participate, and filled out our smoking questionnaire. These smokers were randomly split among three groups:

- *Tailored.* These smokers received the letter generated by STOP from their questionnaire.
- *Non-tailored.* These smokers received the non-tailored letter (Section 3.2).
- *No-letter.* These smokers just received a letter thanking them for participating in our study.

After six months we sent a followup questionnaire asking participants if they had quit, and also other questions (for example, if they were intending to try to quit even if they had not actually done so yet). Smokers could also make free-text comments about the letter they received. 2045 smokers responded to the followup questionnaire, of which 154 claimed to have quit. Because people do not always tell the truth about their smoking habits, we asked these 154 people to give saliva samples, which were tested in a lab for nicotine residues. 99 smokers gave such samples, and 89 of these were confirmed as non-smokers.

The clinical trial took 20 months to run (of which the first 4 months overlapped software development), and cost about UK£75,000 (US$110,000). We believe the STOP clinical trial was the longest and costliest evaluation ever done of an NLG system. The length and cost of the clinical trial were primarily due to the large numbers of subjects. Whereas most previous task-effectiveness evaluations of NLG systems [7,10,13,38] used fewer than 50 subjects, we had 2553 subjects in our clinical trial. The reason the trial needed to be so large was that we were measuring a binary outcome variable (laboratory-verified smoking cessation) with a very low positive rate (since smoking is a very difficult habit to quit).

4.2. Results

Of the 2553 smokers in the trial, 89 were validated as having stopped smoking. These broke down by group as follows:

- 3.5% (30 out of 857) of the tailored group stopped smoking.
- 4.4% (37 out of 846) of the non-tailored group stopped smoking.
- 2.6% (22 out of 850) of the no-letter group stopped smoking.

The non-tailored group had the lowest number of heavy smokers (more than 20 cigarettes per day), who are less likely to stop smoking than light smokers (because they are probably addicted to nicotine); the tailored group had the highest number of heavy smokers. After adjusting for this fact, cessation rates were still higher in the non-tailored group than in the tailored group, but this difference was not statistically significant. Our overall conclusion was therefore that recipients of the non-tailored letters were more likely to stop than people who got no letter (significant at $p = 0.069$ overall and $p = 0.049$ for light smokers).

However, there was no evidence that the tailored letters were any better than the non-tailored ones in terms of increasing cessation rates.

There is some very weak evidence that the tailored letter may have been better than the non-tailored letter among smokers for whom quitting was especially difficult. For example, among discouraged smokers (people who wanted to quit but were not intending to quit, usually because they did not think they could quit), cessation rates were 60% higher among recipients of tailored letters than recipients of non-tailored letters, but the numbers were too small to reach statistical significance, since very few such people managed to stop smoking. Among heavy smokers, again too few people quit to allow statistically significant conclusions about cessation rates, but recipients of the tailored letter were 50% more likely than recipients of the non-tailored letters to show increased intention to quit (for example, say in their initial questionnaire that they did not intend to quit, but say in the followup questionnaire that they did intend to quit) (significant at $p = 0.059$). It would be nice to test the hypothesis that tailored letters were effective among discouraged smokers or heavy smokers by running another clinical trial, but such a trial would need to be even bigger and more expensive than the STOP trial, in order to have enough validated quitters from these categories to make it possible to draw statistically significant conclusions.

Recipients of the tailored letters were more likely than recipients of non-tailored letters to remember receiving the letter (67% vs. 44%, significant at $p < 0.01$), to have kept the letter (30% vs. 19%, significant at $p < 0.01$), and to make a free-text comment about the letter (20% vs. 12%, significant at $p < 0.01$). However, there was no statistically significant difference in perceptions of the usefulness and relevance of the tailored and non-tailored letters.

Free-text comments on the tailored letters were varied, ranging from *I carried mine with me all the time and looked at it whenever I felt like giving in* to *I found it patronising . . . Smoking obviously impairs my physical health—not my intelligence!* The most common complaint about content was that not enough information was given about practical 'how-to-stop-smoking' techniques. In fact the rules in STOP which decided whether to include such content were largely based on the Stages of Change [26] theoretical model of behaviour change; in retrospect it probably would have been better to rely on our KA activities for these rules, as we did in most other cases. Note that since all recipients of the non-tailored letter received 'how-to-quit' advice, the non-tailored letters may have been more effective in this regard.

As a final note, while the absolute cessation rates in STOP were much lower than the cessation rates cited in some other studies of tailored smoking cessation material, this is largely because of different populations and measurement techniques. For example, Strecher et al. [36] reported a 19% cessation rate among recipients of tailored letters in their Experiment 2, which sounds much higher than the 4.4% cessation rate for our non-tailored group. However, Strecher et al. worked with a population of relatively-likely-to-stop smokers (light smokers who were intending to quit); omitted smokers who failed to respond to the followup questionnaire from their percentage calculations; and did not laboratory-validate cessation claims. The cessation rate among recipients of STOP non-tailored letters when calculated in this fashion is 22%, which in fact is higher than the 19% reported by Strecher et al. for recipients of their tailored letters.

4.3. Why did STOP fail

The STOP clinical trial did not tell us why STOP failed. There are many possible reasons for the negative result, including:

(1) Tailoring cannot have much effect in this kind of application. That is, if a smoker receives a letter from his/her doctor about smoking, then the content of the letter is only of secondary importance, what is most important is simply the fact of having received a communication from his/her doctor encouraging smoking cessation.
(2) Tailoring could have an impact, but only if it was based on much more knowledge about the smoker's circumstances than is available via a 4-page multiple choice questionnaire.
(3) Tailoring based on a multiple-choice questionnaire can work, we just did not do it right in STOP, due to inappropriate NLG or KA techniques, and/or invalid theoretical models of behaviour change.
(4) The STOP letters did in fact have an effect on some groups (such as heavy or discouraged smokers), but the clinical trial was too small to provide statistically significant evidence of this.

In other words, did we fail because (1) what we were attempting could not work; (2) what we were attempting could only work if we had a lot more knowledge available to us; or (3) we built a poor system? Or (4) did the system actually work to some degree, but the evaluation failed to show this because it was too small? This is a key question for NLG researchers and developers (as opposed to doctors and health administrators who just want to know if they should use STOP as a black-box system), but the clinical trial does not distinguish between these possibilities.

Plausible arguments can be made for all of the above possibilities. For example, we could argue for (1) on the basis that brief discussions about smoking with a doctor have about a 2% success rate [17], and this may be an upper limit for the effectiveness of a brief letter from a doctor. If so, then letters cannot do much better that the 1.8% increase in cessation rates produced by the STOP non-tailored letter. Or we could argue for (2) by noting that during a number of informal discussions with smokers about our project, many people commented that in-depth knowledge of their circumstances would lead to more effective tailoring. We could also argue for (3) because there clearly are many ways in which the tailored letters could have been improved; and for (4) on the basis of the weak evidence for this mentioned in Section 4.2.

The bottom line is that we cannot draw any firm conclusions from the clinical trial as to what the underlying reason(s) were for STOP's failure. If STOP had worked, then we could have concluded that (1) tailoring can work, (2) tailoring does not require in-depth knowledge of the user, (3) STOP did a good job of tailoring, and (4) the STOP clinical trial was effective at demonstrating this. Since STOP did not work, we know that at least one of these hypotheses must be false, but we do not know which one(s).

5. General discussion: Negative results

5.1. Negative results in science

Negative results are usually held to be an essential part of science. For example, the philosopher of science Karl Popper [24] argued that *falsification* and *falsifiability* were at the heart of scientific research; in other words, that experimental science should focus on attempts to disprove theories. Popper argued that it was impossible to experimentally prove theories, because scientists cannot experimentally check all predictions of a theory in all possible contexts; what experimental science can do is disprove theories, by showing instances where a theory makes false predictions. Popper also stated that negative results were the main motivation for theorists searching for better theories [24, p. 108]; in other words, negative results are the main driving force behind improvements in scientific theory.

However, negative results that clearly disprove theories in Popper's sense may be rare, except in retrospect. In particular, it is often unclear except in retrospect whether a negative result falsifies an entire theory, requires a theory to be tweaked in a relatively minor way, or is simply due to bad experimental design. For example, Popper cites (p. 108) the Michelson–Morley experiment as an example of an experiment falsifying a theory (in this case that light travelled through an 'ether'), and hence leading to a better theory (Einstein's relativity). However, at the time it was performed the Michelson–Morley experiment was not perceived of as falsifying the theory of ether, but rather as requiring the theory to be tweaked, for example by assuming that the Earth dragged an envelope of ether around it [16, pp. 29–30]. It was only 18 years later, when Einstein published his theory of relativity, that the Michelson–Morley experiment began to be seen as a possible falsification of the whole theory of ether.

To take a more recent example which is closer to STOP, Aveyard et al. [1] reported that an interactive multimedia 'expert system' was not effective at smoking cessation and prevention in secondary schools. The developers of the expert system argued that this negative result was due to methodological problems with the experiment [25], but Aveyard et al. disputed this [2]. So in this case, as in many cases in medicine, the negative result is clear, but there is debate as to what exactly it means.

Perhaps because of the difficulty of drawing clear lessons from them, negative results are probably published less often than positive results in the scientific community. The medical community in particular is concerned about this, because:

- it is unethical to subject patients to a clinical trial of an experimental technique that previous experiments have shown not to work, especially if this means denying patients access to a control technique which is known to work; and
- meta-analyses, which look for patterns in and otherwise analyse groups of related experiments, will give incorrect results if the only experiments included in the analysis are those that gave positive results.

Dickersin and her colleagues [11,12,22] have analysed the failure to publish negative results in medicine in several studies. She has concluded that the main problem is the reluctance of researchers to submit papers with negative results; her work suggests

that once submitted, papers presenting negative results are as likely to be accepted and published as papers presenting positive results. Dickersin [11] further suggests that the reluctance of researchers to publish negative results is largely because researchers themselves do not regard negative results as being as exciting as positive results, so they do not place a high priority on publishing them. In other words, if a researcher has a limited amount of time for writing up experimental results, he or she is likely to write up the positive results first, and the negative ones only later, if time permits. However, despite this many medical papers with negative results are still published; a study performed at the *Journal of the American Medical Association* [22] on papers describing randomised controlled clinical trials found that 46% of submitted papers and 38% of accepted papers gave negative results.

5.2. Negative results in artificial intelligence

Negative results have historically been rare in AI, perhaps because until the 1990s most AI researchers did not experimentally test hypotheses and indeed often did not even pose hypotheses that were specific enough to be falsifiable (testable). When AI 'research' consisted of building complex programs and demonstrating these programs on a few hand-picked examples, negative results were almost unthinkable (if a program did not work quite as expected on an example, the researcher would often just tweak the program or look for another example). Cohen [9, p. xii] surveyed 150 papers in the *Proceedings of the Eighth National Conference on Artificial Intelligence* (*AAAI-1990*), and found that only 21% of these papers even presented hypotheses or predictions, and that 'very few' papers reported negative or unexpected results.

More recently, though, there has been a much greater emphasis on experimentation in AI in general and NLP in particular. Given this, one would expect to see more papers describing negative results, but such papers still seem few and far between.

For example, of the 68 (non-invited) papers presented at the 2001 conference of the Association for Computational Linguistics (ACL), 32 (47%) presented experimental results which compared a proposed algorithm, model or system against a control. This is a reasonable percentage, given that ACL also includes theoretical papers, papers on tools and resources, papers on novel systems for which there is no clear control for experimental comparison, etc. However, only 4 of the experimental papers presented negative results, and 2 of these papers seemed to regard the negative result as primarily signalling that programs or algorithms needed to be further developed. Only 2 of the 32 experimental papers (6%) [14,31] (including one paper about STOP) presented negative results as scientifically meaningful findings that potentially cast doubt on hypotheses.

5.3. Why so few negative results in AI?

There are many possible reasons for the paucity of negative result papers in AI as compared to medicine. One is that negative results in AI are indeed rare, and the great majority of experimental tests of AI systems and hypotheses are positive. If true, however, this suggests that AI researchers are either risk-averse or incompetent (or dishonest) at experimental evaluation. Just as in any other field of human endeavour, a 94% success

rate suggests either reluctance to try anything risky or challenging, or cheating. Lack of negative results would also suggest that AI researchers are not scientists in Popper's sense, since they are reluctant to falsify theories.

Another possibility is that editors and peer reviewers refuse to accept papers reporting negative results. We have no bibliometric evidence about this, of the sort collected by Dickersin, but our personal experiences as paper authors and reviewers suggest that this is untrue. We strongly suspect that the situation is analogous to the one reported by Dickersin in medicine; editors and reviewers are open to papers reporting negative results, but few such papers are in fact submitted to AI journals and conferences, because AI researchers are even more reluctant than medical researchers to submit papers giving negative results.

We wonder if this reluctance is in part because the scientific value and most appropriate presentation of negative results in AI is less clear than in medicine. A negative result which clearly falsifies a major theory in Popper's sense is clearly worth reporting, but as mentioned above, most negative results probably do not lead to such unambiguous interpretations. They indicate that something is wrong with the experiment or the theory, but they may not pinpoint exactly where the problem is. The medical community knows that negative results are worth reporting regardless (unless they have major experimental flaws), because they provide essential data for meta-analyses and because not reporting negative results may lead other scientists to conduct unnecessary clinical trials. But AI has much less experience with negative results and many fewer role models, which may discourage the submission of papers which cannot present clear lessons from a negative result.

We believe that in principle negative results are important in AI for much the same reasons as they are important in medicine; not just direct theory falsification, but also as essential data for theory formation, and as warnings to other researchers that certain research paths may not be productive. Certainly we know from machine learning research that any attempt to learn a theory from data is much more likely to work if the data contains negative as well as positive examples. And while AI researchers do not generally have to worry about the ethical implications of clinical trials, we do think that AI researchers have a duty to inform their colleagues if they have evidence that a research direction may not be fruitful. We have a colleague who wasted a considerable amount of time during his PhD because certain negative results were not published, and we suspect he is far from unique.

We believe that many of our colleagues share these sentiments, at least in abstract. However, unlike medical researchers, we AI researchers do not have a large set of examples, role models, and 'case law' on when negative results should be published, how they should be described, and indeed on how experiments should be organised in order to maximise the scientific usefulness of negative as well as positive outcomes.

5.4. What can be learned from STOP?

So, what can the AI community learn from the specific negative result discussed in this paper?

As discussed in Section 4.3, STOP did not unambiguously falsify any hypotheses, except the uninteresting (to a computer scientist) hypothesis that STOP was medically effective at reducing smoking rates; and even here it is in fact possible that the system was effective

with some groups of smokers but the clinical trial was too small to show this. As discussed above, such ambiguity is not uncommon; a negative result indicates that there is a flaw somewhere in theory or experiment, but it may not indicate exactly where the flaw lies.

However, we hope that STOP will provide a useful data point for people interested in forming theories about NLG, knowledge acquisition, user adaptation, and AI in medicine. For example, the clinical trials of both STOP and PIGLIT [8] showed no effect on patient behaviour, but the PIGLIT trial showed an effect on patient subjective satisfaction. Could it be that AI patient-information systems are a good way of making patients happier with their health care but should not be expected to change patient behaviour? Of course we need many more data points (that is, experimental trials of AI patient-information systems) before we can make such statements with any confidence; but this example perhaps illustrates the potential power of meta-analyses of related experiments and projects.

We also hope that our lack of success in STOP will be taken into consideration by other people who are considering similar projects. This is not of course to say that a STOP-like system could not work, but we would advise that anyone who is considering building such a system should think about how their planned effort differs from STOP, and how likely these differences are to lead to a more successful outcome.

Finally, we believe that STOP has highlighted a few places where better AI theories and methodologies are needed. For example, we need better knowledge acquisition techniques for tasks that experts do not currently perform (such as writing tailored smoking cessation letters); and a better understanding of the ways in which generated texts can be tailored (for example, positive vs. negative tone) and the impact of such tailoring on recipients. We also need a better understanding of large-scale evaluations of AI systems with real subjects or users: how should such evaluations be designed, what can we expect to learn from them, and in general when are such evaluations appropriate and when should cheaper techniques be used?

6. Conclusion

STOP was an NLG system which generated tailored stop-smoking letters. Perhaps unusually for an AI system, it was rigorously evaluated for effectiveness with substantial numbers of end users (smokers). Unfortunately, this evaluation showed that STOP was not in fact effective at meeting the goal of increasing smoking cessation rates. We cannot pinpoint the reason for this failure; possibilities include lack of sufficient knowledge about smokers, inadequate KA or NLG techniques, an impossible application, and a too-small evaluation. This means that we cannot make a clear statement about what STOP did and did not prove. Nevertheless, we believe that STOP is useful as a case study, and that it raises questions that the AI community should consider. In particular, how (and when) should negative results be published, and what should other AI researchers expect to learn from negative results?

Our experience in STOP was that the medical researchers involved in the project had a much clearer idea of what to do with negative results than the AI researchers, and we believe that this reflects the greater maturity of medicine as a field of empirical scientific research. We hope that our paper will encourage the AI community to think more about

how, when, and why negative results should be reported, because we believe that knowing what to do with negative results is an essential aspect of empirical scientific research.

Acknowledgements

Many thanks to the rest of the STOP team, including Scott Lennox, Yllias Chali, Peter Donnan, James Friend, Annette Hermse, Duncan MacIver, Ian McCann, Yvonne McKay, Steven Porter, and Diane Skatun. Thanks also to the many people who commented on this paper and on earlier conference papers about STOP. This research was supported by the Scottish Office Department of Health under grant K/OPR/2/2/D318, and by the Engineering and Physical Sciences Research Council under grant GR/L48812.

References

[1] P. Aveyard, K. Cheng, J. Almond, E. Sherratt, R. Lancashire, T. Lawrence, C. Griffin, O. Evans, Cluster randomised controlled trial of expert system based on the Transtheoretical (Stages of Change) model for smoking prevention and cessation in schools, British Medical J. 319 (1999) 948–953.
[2] P. Aveyard, K. Cheng, T. Lawrence, Reply to Prochaska et al. letter (2000), http://bmj.com/cgi/content/full/320/7232/447.
[3] S. Bangalore, O. Rambow, S. Whittaker, Evaluation metrics for generation, in: Proceedings of the First International Conference on Natural Language Generation, Mitzpe Ramon, Israel, 2000, pp. 1–8.
[4] D. Bental, A. Cawsey, R. Jones, Patient information systems that tailor to the individual, Patient Education and Counselling 36 (1999) 171–180.
[5] K. Binstead, A. Cawsey, R. Jones, Generated personalised patient information using the medical record, in: P. Barahona, M. Stefanelli, J. Wyatt (Eds.), Proceedings of the Fifth Conference on Artificial Intelligence and Medicine Europe (AIME-1995), Springer, Berlin, 1995, pp. 29–41.
[6] B. Buchanan, D. Wilkins (Eds.), Readings in Knowledge Acquisition and Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[7] G. Carenini, J. Moore, An empirical study of the influence of argument conciseness on argument effectiveness, in: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, 2000, pp. 150–157.
[8] A. Cawsey, R. Jones, J. Pearson, K. Binstead, The design and evaluation of a personalised health information system for patients with cancer, User Modelling and User-Adapted Interaction 10 (1999) 47–72.
[9] P. Cohen, Empirical Methods for Artificial Intelligence, MIT Press, Cambridge, MA, 1995.
[10] N. Colineau, C. Paris, K. Vander Linden, An evaluation of procedural instructional text, in: Proceedings of the Second International Conference on Natural Language Generation (INLG-2002), New York, 2002, pp. 128–135.
[11] K. Dickersin, Y. Min, NIH clinical trials and publication bias, Online J. Current Clinical Trials 50 (1993).
[12] K. Dickersin, Y. Min, C. Meinert, Factors influencing publication of research results, J. Amer. Medical Assoc. 267 (1992) 374–378.
[13] B. Di Eugenio, M. Glass, M. Trolio, The DIAG experiments: Natural language generation for tutoring systems, in: Proceedings of the Second International Conference on Natural Language Generation (INLG-2002), New York, 2002, pp. 120–127.
[14] M. Johnston, Joint and conditional estimation of tagging and parsing models, in: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001), Toulouse, France, 2001, pp. 314–321.
[15] R. Jones, J. Pearson, S. McGregor, A. Cawsey, A. Barrett, N. Craig, J. Atkinson, W. Gilmour, J. McEwen, Randomised trial of personalised computer based information for cancer patients, British Medical J. 319 (1999) 1241–1247.

[16] D. Kevles, The Physicists, Knopf, 1977.
[17] M. Law, J. Tang, An analysis of the effectiveness of interventions intended to help people stop smoking, Archives of Internal Medicine 155 (1995) 1933–1941.
[18] S. Lennox, L. Osman, E. Reiter, R. Robertson, J. Friend, I. McCann, D. Skatun, P. Donnan, The cost-effectiveness of computer-tailored and non-tailored smoking cessation letters in general practice: A randomised controlled study, British Medical J. 322 (2001) 1396.
[19] J. Levine, C. Mellish, The IDAS user trials: Quantitative evaluation of an applied natural language generation system, in: Proceedings of the Fifth European Workshop on Natural Language Generation, Leiden, The Netherlands, 1995, pp. 75–93.
[20] D. McDermott, Artificial intelligence meets natural stupidity, in: J. Haugland (Ed.), Mind Design, MIT Press, Cambridge, MA, 1981, pp. 143–160.
[21] C. Mellish, R. Dale, Evaluation in the context of natural language generation, Computer Speech and Language 12 (1998) 349–373.
[22] C. Olson, D. Rennie, D. Cook, K. Dickersin, A. Flanagin, J. Hogan, Q. Zhu, J. Reiling, B. Pace, Publication bias in editorial decision making, J. Amer. Medical Assoc. 287 (2002) 2825–2828.
[23] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002). Philadelphia, PA, 2002, pp. 311–318.
[24] K. Popper, The Logic of Scientific Discovery, Hutchinson and Co, London, 1959, translated from German by the author.
[25] J. Prochaska, Stages of change model for smoking prevention and cessation in schools (letter to editor), British Medical J. 320 (2000) 447.
[26] J. Prochaska, C. diClemente, Stages of Change in the Modification of Problem Behaviors, Sage, 1992.
[27] E. Reiter, Shallow vs deep techniques for handling linguistic constraints and optimizations, in: Proceedings of the KI-1999 Workshop on May I Speak Freely: Between Templates and Free Choice in Natural Language Generation, 1999.
[28] E. Reiter, Pipelines and size constraints, Computational Linguistics 26 (2) (2000) 251–259.
[29] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, 2000.
[30] E. Reiter, R. Robertson, The architecture of the STOP system, in: Proceedings of the Workshop on Reference Architectures for Natural Language Generation, Edinburgh, UK, 1999.
[31] E. Reiter, R. Robertson, S. Lennox, L. Osman, Using a randomised controlled clinical trial to evaluate an NLG system, in: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001),Toulouse, France, 2001, pp. 434–441.
[32] E. Reiter, R. Robertson, L. Osman, Knowledge acquisition for natural language generation, in: Proceedings of the First International Conference on Natural Language Generation, Mitzpe Ramon, Israel, 2000, pp. 215–217.
[33] E. Riloff, An empirical study of automated dictionary construction for information extraction in three domains, Artificial Intelligence 85 (1996) 101–134.
[34] A.C. Scott, J. Clayton, E. Gibson, A Practical Guide to Knowledge Acquisition, Addison-Wesley, Reading, MA, 1991.
[35] V. Strecher, Computer-tailored smoking cessation materials: A review and discussion, Patient Education and Counselling 36 (1999) 107–117.
[36] V. Strecher, M. Kreuter, D.-J. Den Boer, S. Kobrin, H. Hospers, C. Skinner, The effects of computer-tailored smoking cessation messages in family practice settings, J. Family Practice 39 (1994) 262–271.
[37] M. White, T. Caldwell, EXEMPLARS: A practical extensible framework for dynamic text generation, in: Proceedings of the Ninth International Workshop on Natural Language Generation (INLG-1998), Niagara-on-the-Lake, ON, 1998, pp. 266–275.
[38] M. Young, Using Grice's maxim of quantity to select the content of plan descriptions, Artificial Intelligence 115 (1999) 215–256.