Evaluating a Natural Language Dialogue System: Results and Experiences

**Nick Webb[†], Anne De Roeck, Udo Kruschwitz,
Paul Scott, Sam Steel and Ray Turner**

Department of Computer Science
University of Essex, Colchester, UK
yp@essex.ac.uk

Abstract

Dialogue Management components are becoming increasingly useful in Natural Language Information Retrieval systems, to aid users find information they require. These components need to be evaluated, both intrinsically, to gauge development approaches, and extrinsically to measure subjective performance. We introduce a Natural Language Dialogue System, the YPA, developed at the University of Essex, that accesses classified directory information such as British Telecom's Yellow Pages. These directories contain much useful but hard-to-access information, especially in the free text in semi-display advertisements. We examine the evaluation methodologies used to assess the performance of the YPA during development, both from users and technology developer perspectives, by using both subjective and objective evaluation approaches. We finish by drawing some conclusions from these experiments.

1. Introduction

The increase in the volume of data available to users has made simple question-answering systems almost defunct. So much data can match even simple queries that systems now have to be able to determine the exact information that users require. To this end, Dialogue Management (DM) components are becoming a necessity in Natural Language Information Retrieval. These components try to resolve ambiguity between user requirements and available data, either through using some domain knowledge source, interaction with the user or a combination of both.

The evaluation of these Natural Language Dialogue systems (NLDs) is a complex task. It has become important to be able to demonstrate the increased effectiveness of the DM component, and of the NLDs as a whole. There are two reasons for doing this.

Firstly, intrinsic evaluation is of vital importance to system designers, as a measure of success or otherwise of their current work. More recently, the notion of extrinsic evaluation has become important, where it is necessary to demonstrate an increase of performance in relation to some other system.

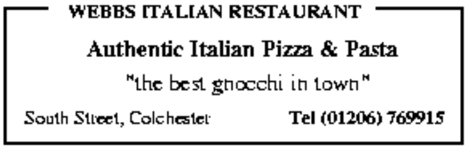
This paper deals specifically with the first of these two goals, and details the three evaluations that have been performed on a NLD system, highlighting objectives and usefulness of results. In section 2 we introduce the YPA, a directory enquiry system which allows a user to access advertiser information in classified directories (De Roeck et al., 2000) and in section 3 describe evaluations performed on the YPA so far. We draw conclusions from this in section 4.

2. The YPA

The YPA is an interactive system, which uses a natural language dialogue interface to Talking Pages and Yellow Pages data[1].

Yellow Pages data consists of headings (restaurants, plumbers…), advertiser name, contact information, and often some form of 'free text', where advertisers can write something about themselves or their products ("best gnocchi in town", as in figure 1 for example). Talking Pages data is much the same, with far more free text per advert, sometimes as much as 5 pages of data.

We wanted a system that would allow users to enter exactly that information they required, independent of any indexing scheme employed on the data. For example, if the user requires gnocchi, traditionally they could look under restaurants, or under take-away food, but in either case, they will have to read each advertisement to discover if they serve gnocchi, or more likely if the restaurant serves Italian food. The YPA allows a user to search over the content of advertisements, rather than manually assigned keywords and business classifications, relieving the need to know which classification the desired advertisement will appear under. The method of extracting keywords from the advertisements is described in (Kruschwitz et al., 2000).



WEBBS ITALIAN RESTAURANT

Authentic Italian Pizza & Pasta

"the best gnocchi in town"

South Street, Colchester Tel (01206) 769915

Figure 1: Simple advertisement with free text

The aim for the system was to have a high precision of returned results. If there was a match for a query in the database, we wanted to be sure we could find it.

A conversation cycle with the YPA can be roughly described as follows. A user utterance (typed in via the Graphical User Interface) is sent to the Dialogue Manager. The Dialogue Manager keeps track of the current stage in the dialogue and controls the use of several sub-modules. Before handing back control (together with the relevant data) to the Toplevel, the input is first sent to the Natural

[†] Nick Webb is now a member of the Natural Language Processing Group at the University of Sheffield, UK. n.webb@sheffield.ac.uk

[1] Yellow Pages® and Talking Pages® are registered trademarks of British Telecommunications plc in the United Kingdom.

Language Frontend which returns a so-called slot-and-filler query. The Dialogue Manager then consults the Query Construction Component, passing it the result of the parsing process (possibly modified depending on the dialogue history, for instance). The purpose of the Query Construction Component is to transform the input into a database query (making use of the Backend and possibly the World Model), to query the Backend and to return the retrieved addresses (and some database information) to the Dialogue Manager. Finally the Dialogue Manager hands back control to the Toplevel which for example displays the retrieved addresses. It could also put questions to the user that were passed to it by the Dialogue Manager, if the database access was not successful (i.e. did not result in a set of addresses). At this stage the cycle starts again.

3. Evaluation

Evaluating any NLD system can take either an objective (possibly using precision, recall or some combination of both) or subjective (user-satisfaction) approach. In order to have the best guide to the development of the YPA we have performed both. In addition, all evaluations were performed on the system as a whole, rather than attempting to target individual components. It can be difficult to assess the performance of individual components of systems such as these (see Dybkjær et al., 1998), because the functionality of individual modules is often dependent on other modules within the system.

For preliminary evaluation we used the YPA for the Colchester area Yellow Pages, with about 26,000 addresses (De Roeck et al., 1998).

The technology-focused approach was conducted in September of 1998. It was the objective to see if the system could process queries successfully and identify initial problems at this early stage. Also it presented our first opportunity to collect a corpus of queries directed at a Yellow Pages system.

The second evaluation, in December 1998, was also a mostly objective study. We attempted to measure the precision of the recalled addresses with respect to the query, but this also involved a degree of subjectivity - what constituted a relevant address. In August 1999 we had the most recent evaluation, which is far more subjective. We measured users reactions to the system, and recorded their impression of the accuracy of the answers retrieved. It also differed from the previous evaluations in that we used the YPA loaded with Talking Pages data covering parts of the Bristol and London areas (consisting of about 13,000 addresses). Talking Pages data is much more detailed in the amount of free text which relates to each advertiser (De Roeck et. al. 2000).

We examine each of these evaluations in further detail.

3.1. Technology-Focused Evaluation

During a two week period, members of British Telecom's Intelligent Systems Research Group had access to the YPA via the World Wide Web. They were given an outline of the information the system contained, but not told how the system should be queried. Over this time the administrators of the system collected some 238 unique queries. These queries were used to conduct the technology-focused evaluation of the YPA system. Each query was evaluated to see if it presented a list of addresses to the user, either first time or as the result of some dialogue with the system. If a list of addresses was not returned, we recorded the reason for this, and presented the figures as percentages of the total number of queries.

Of these queries, some 68% were deemed to be a system success, where the system performed as it was designed to, taking no account of quality of results or potential user satisfaction.

Of the queries that failed, 10% were due to some failure to correctly process the input (mainly a problem with our Natural Language Frontend, or the stemmer). A further 12% failed due to a problem with our World Model – caused during the query expansion phase. For example, the query:

'I want to buy a car magazine'

cannot be matched exactly to the data we have in the database. This causes the Query Construction Component to try to expand the query, by using WordNet (Miller 1990) synonym and hypernym relations. At this stage in the project, such expansion occurred in parallel, i.e. all terms that could be expanded would be expanded, and this in turn would lead to:

- car ⇒ machine
- magazine ⇒ product

and the resulting new search for a *'machine product'* would produce addresses completely irrelevant to the initial query.

The final 10% of problems were caused by users - some typing errors, some people using out-of-bounds locations and some making statements of fact rather than asking queries (so we are less able to cope with input of the type *'I am dying for a beer'* – which returned a list of funeral services).

We took the results of this evaluation to conclude that the system was performing as it was designed to, and that the robustness of the system was evident in that there were no fatal hardware or software errors during the time the system was online. We managed to collect a useful corpus of Yellow Pages queries, which could be used to guide the development of a new version of the system.

Furthermore, we could use the results of this evaluation to produce figures for the performance of the parsing strategy employed by our Natural Language Frontend (Webb et al., 1999).

The evaluation of parsing strategies used in Natural Language Frontend construction is an art in itself. There have been a large number of suggested methods for the purpose of guiding and monitoring the development of a parsing system (see Carroll et al., 1998 for a good survey).

We used the method of coverage, where we calculate the percentage of sentences assigned an analysis from an unannotated corpus. This method of evaluation is very crude, given that the rules in the grammar allow a variety of parse trees, which might not necessarily be correct. Although linguistic correctness was not a requirement, it is equally important that we have some form of correct structural identification, as we aim to preserve some structural information throughout the slot-filling procedure.

90% were successful from the parser point of view. 10% were recorded a Frontend failure - where either queries failed the parse completely, or information was misplaced, and put in incorrect slots.

Once a new system had been created in light of the findings of the technology evaluation the next obvious step then was to measure the performance of the system as a whole, to test our claim that our system returned a high degree of precision for input queries.

3.2. Precision-Focused Evaluation

At this stage, we wanted a measure of how good the results produced by the YPA were (and we knew from the first evaluation that 90% of the queries produced some sort of results). In fact we wanted to measure how closely retrieved addresses matched the input query. Importantly this takes no account of how satisfied prospective users would be with these answers, nor whether the system produced all available addresses. However, we did presume there would be some sort of correlation between the number of relevant addresses and user satisfaction.

We developed a response sheet for each query, recording:

- whether addresses were offered or not, and if not, why?
- how many dialogue turns it took to get addresses
- if any system initiated dialogue seemed relevant?
- total number of addresses recalled
- number of recalled addresses judged to be relevant

The last two in this list would allow us to calculate a percentage of those addresses that were relevant.

For this evaluation, we used a corpus of 75 queries, collected from the Talking Pages call centre in Bristol. The only alteration was to change any specified location information to match an appropriate area in our data, i.e. Colchester, or some area therein.

Of these 75 queries, 62 (83%) asked for information about addresses contained within our sample set (that is, existed within the Colchester area). For those, we had a value of 74% of addresses returned that were relevant to the input query.

The issue of how to deal with 'negative return queries', that is how to satisfactorily inform the user that no answers can be found is probably a major factor in user satisfaction with any system. The problem can be highlighted using a couple of simple examples.

'I want the number of barclaycard gold'

The system returns by saying that there is nothing in the database about *barclaycard gold*, a fair indication that this advertiser does not exist in the database.

'Dux's museum'

The system returns with a list of addresses of museums, but informs the user that none match all of the query terms (as Dux's museum does not exist in this data set). A good attempt to provide alternatives, but if the user was after a specific museum, would they be happy with this?

In terms of our overall precision statistic, how should these queries with negative returns be scored? Clearly the questions posed by the user have been answered to some degree, but assigning these queries a value of 100% each seems unrealistic. Instead, we felt it was better to create for ourselves a worst case scenario, where we would assign any query with a negative return a 0% precision value. In terms of our evaluation, this meant assigning the 13 queries in our corpus that would have negative returns a value of 0%. Combining this with our 74% precision figure would give us a base-line figure, which in this evaluation was 61%.

We clearly felt that the system was actually performing above this level, but in order to show that we needed to conduct some sort of user satisfaction survey, to gauge how users felt about the answers returned by the system.

3.3. User Satisfaction

At the end of the YPA project, we will have to perform some full user tests. Before that event however, we needed to have some preliminary user trials, to judge the performance of the system, and to evaluate our evaluation procedure. We traveled to the Talking Pages call centre, and enlisted the help of five Talking Pages operators.

The current system in use at the Talking Pages call centre is based on keyword searches, with keywords being assigned to individual advertisements based on the content, but not necessarily drawn from it. Regularly occurring features could be indexed (taking Visa or MasterCard), but not other, less obvious regularly occurring parts of free text. Problems with this system occurred when both users and operators were unable to determine which business classification to look under – the first step required by the system.

We asked the operators to query our system, first of all using a set of queries we provided (which had been selected from tapes of Talking Pages queries supplied to us previously) then using any queries they provided themselves. We asked them to place specific emphasis on any queries that the current Talking Pages system found difficult.

For each query, an evaluation form was completed (called the last query form) and once each user had completed their time with the system they filled out a form to capture their feelings about the system as a whole (the overall impression form).

3.3.1. Last Query Form

The last query form was divided into two sections, depending on whether or not the query produced a list of addresses. If addresses were produced, then users were asked to rate those addresses in terms of relevance on a Likert scale, from 1 (very poor) to 5 (very good). Then they were asked four yes/no questions:

- Was there an address that satisfied the query?
- Were the addresses presented in an acceptable order?
- Did the system ask you any questions?
- Were the questions asked by the system sensible?

If the query failed to produce a list of addresses, the user was led to a different set of three yes/no questions:

- Did the dialogue reach a satisfactory conclusion?
- Do you believe that the information is in there but you cannot get at it?
- Were the questions asked by the system sensible?

In addition to these, we provided the user with ample space to make comments on the performance of the system with regard to the specific query.

We collected 33 of these forms in the course of our evaluation. Clearly such a small sample set is not representative of system performance, but we will use it as an indicator for further evaluation. Taking all the queries, 32 of the 33 produced a list of addresses, and of those users judged the relevance of the answers to have an average value of 3.97.

Other results looked like this:

- In 72% of cases users felt they got an answer to their query;
- 78% of the time the addresses were presented in an acceptable order;
- 50% of queries caused the system to initiate some dialogue - and of those, 94% of the dialogue steps presented were judged to be sensible.

Only one query did not retrieve any addresses, and in this instance the user judged that the information could not be found in the database, that the dialogue reached a satisfactory conclusion and that the questions asked by the system were sensible.

3.3.2. Overall Impression Form

The Overall Impression Form was designed similarly to that used in the Eurospeech `97 ELSNET Olympics, described in (den Os and Bloothooft, 1998). It consisted of five system aspects to be scored on a Likert scale identical to that of the last query form, which would measure:

- Overall performance of the YPA
- Relevance of addresses found
- Order of displayed addresses
- Length of interaction
- Appropriateness of responses in the interaction

There were only three of these forms completed, making the statistical results of little value, but they are included here for completeness. Overall performance was rated as a 4.0 (all users indicated this). The relevance of address was 3.7 (similar to that calculated from the last query forms). The order of displayed addresses was appraised as good with a score of 4.0, and both the length and appropriateness of interaction were rated at 4.3.

Further to these there were three text boxes. One asked the user to identify specific areas of the system they felt were strong, one to highlight weak points and finally a general comments section.

3.3.3. User Comments

Nothing has yet been said about the entries in the free text fields on either of our forms, yet this is probably the most interesting bit both to system developers and to those with an interest in using the system.

Positive comments focused on the need in the Talking Pages system to have extensive prior knowledge of the indexing system, something not necessary with the YPA, and that the YPA was 'very precise' in its answers. Additionally, users were impressed by the speed of the system in retrieving addresses, especially pleasing when considering the whole trial was conducted over a dial-up Internet connection.

Negative comments concerned aspects of the YPA which were not yet implemented, nor indeed were they part of the research in this project, such as the spell checker, and the ability to truncate words. More seriously there were some problems with the 'negative return' queries highlighted in the precision focused evaluation. Some users were not happy that on asking for something specific (*'clown outfits'*, for example) they were presented by some more general list which failed to answer the question. In our example, with no clown outfits in the database, a list of outfitters would be presented, as opposed to fancy dress shops. This highlights again the need for further work in this area.

4. Conclusions

This series of evaluations shows clearly the progression of the system, and also justifies our claims that it is a more precise engine for accessing classified directory information.

Our feeling that some combination of relevance of addresses and successful handling of negative return queries can gauge user satisfaction would appear to be correct.

Moreover, we can say that the evaluations we have performed in the laboratory seem (taking the indicators of the final evaluation) to be a true account of how the system is operating, possibly because we used an intermediate evaluation, combining some objectivity with some subjectivity. Using objective (technology focused) and subjective (user focused) studies distinctly can produce widely varying results, as demonstrated by the evaluation of the Dutch VIOS system (Haaren et al., 1998).

At all stages the evaluations have proved extremely useful to monitor and direct the progress of the system.

We continue with our evaluations of the YPA. At the moment we are most concerned about a more detailed evaluation, taking into account recall as well as precision values. We will use the experiences of our other evaluations, especially the most recent, to design forms which will allow us to capture user data more precisely.

5. Acknowledgments

This work has been funded by a contract from the ISR group at BT Laboratories, Martlesham Heath, UK. Most of the data was provided by Yellow Pages. We would like to thank those users at the Talking Pages call centre who helped us with our investigations.

6. References

Carroll, J., T. Briscoe, and A. Sanfilippo, 1998. Parser evaluation: a survey and a new proposal. *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 447-454.

De Roeck, A., U. Kruschwitz, P. Scott, S. Steel, R. Turner, and N. Webb, 2000. The YPA – An Assistant for Classified Directory Enquiries. In B. Azvine, N. Azarmi and D. Nauck (eds.), *Intelligent Systems and Soft Computing: Prospects, Tools and Applications.* Lecture Notes in Artificial Intelligence 1804, 245-264.

De Roeck, A., U. Kruschwitz, P. Neal, P. Scott, S. Steel, R. Turner, and N. Webb, 1998. YPA – an intelligent directory enquiry assistant. *BT Technology Journal*, 16(3):145-155.

den Os, E., and G. Bloothooft, 1998. Evaluating various spoken language dialogue systems with a single questionnaire: Analysis of the ELSNET olympics. *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 51-54.

Dybkjær, L., N. O. Bernsen, R. Carlson, L. Chase, N. Dahlbäck, K. Failenschmid, U. Heid, P. Heisterkamp, A. Jönsson, H. Kamp, I Karlsson, J. v. Kuppevelt, L. Lamel, P. Paroubek, and D. Williams, 1998. The DISC approach to spoken language systems development and evaluation. *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 185-189.

Kruschwitz, U., A. De Roeck, P. Scott, S. Steel, R. Turner and N. Webb, 2000. Extracting Semistructured Data – Lessons Learnt. *Proceedings of the 2nd International Conference on Natural Language Processing (NLP2000)*.

Miller, G. A., 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4) (special issue).

van Haaren, L., M. Blasband, M. Gerritsen, and M. van Schijndel, 1998. Evaluating quality of spoken dialogue systems: Comparing a technology-focused and a user-focused approach. *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 655-660.

Webb, N., A. De Roeck, U. Kruschwitz, P. Scott, S. Steel, and R. Turner, 1999. Natural Language Engineering: Slot-Filling in the YPA. *Proceedings of the Workshop on Natural Language Interfaces, Dialogue and Partner Modelling (at the Fachtagung für Künstliche Intelligenz KI'99).*