Natural Language Engineering **5** (3): 237–250. Printed in the United Kingdom

© 1999 Cambridge University Press

Routing email automatically by purpose not topic

HAMID KHOSRAVI and YORICK WILKS

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
e-mail: {h.khosravi,y.wilks}@dcs.shef.ac.uk

(*Received June 1996; revised May 1999*)

Abstract

In this paper we present results concerning the large scale automatic extraction of pragmatic content from email by a system based on a phrase matching approach to speech act detection combined with empirical detection of speech act patterns in corpora. The results show that most speech acts that occur in this corpus can be recognized by the approach. This investigation is supported by analysis of a corpus consisting of 1000 emails.

1. Introduction

The tradition of routing messages using Information Retrieval (IR) techniques based on key word statistics (Guthrie, Walker and Guthrie 1994) has normally been concerned with content, in the sense of the topic, subject matter or domain relevant to the user. In this paper we describe experimental work to sort a substantial sample of emails based on their function, which is to say, whether they contain a statement of fact, a request for the recipient to do something, or ask a question. We believe this could be a highly desirable functionality for the over burdened email user, especially if combined in use with other, more traditional, measures of relevance and filters based on desirable and undesirable mail sources. Such function-based message routing naturally relates to the tradition of analysis based on speech or dialogue acts, which we describe below, an established philosophic-linguistic manner of describing and detecting such message functions. We shall adopt the overall methodology of Information Extraction (IE) (Gaizauskas and Wilks 1997), a technology developed not to route messages but to extract content directly from text, rather than selecting document subsets as IR does. We have attempted here to apply an IE engine to the extraction of message content located in the message, in part by the use of speech-act detection criteria, e.g. for what it is to be a request for action, under the many possible surface forms that can be used to express that in English. The work may have potential uses in a busy world, but here we describe it as a challenge of adapting an IE engine to a somewhat different, task, that of message function.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

2. Background

Scientists, politicians, academics and businessmen spend much time daily reading, sending and replying to their emails, and it can take hours for busy people to read hundreds of emails, although they know that many of them are probably valueless.

The aim of this paper is to use speech act theory (Searle 1969) to identify the functional gist of variable length emails, which could be used along with other techniques to implement an email filter that will alleviate the problem. Since the response time of our method is reasonably low, it could be used as part of a fast interface between email and deeper analysis of functional content that we make use of here.

There are many possible approaches to speech acts in texts and text categorization (Cohen 1995; Lee and Wilks 1996). Here we make use of an approach that could be called phrase matching, where that is to be taken, as we shall show, in a broader sense than a 'cue-based' approach to speech act detection, such as Hinkelman's (1989).

From a different point of view, Guthrie, Walker and Guthrie (1994) present a theory for determining, for a given document, into which of several categories it best fits. Although they claim that their theory is maximally effective for routing by features, it is not a suitable algorithm for email text processing because:

(i) The theory is based wholly on the frequency of words in text. Email texts, especially in this domain, are too short for significant counts.
(ii) The best practical results they report are achieved for routing between two unrelated subjects: business and terrorism. In the email task, the topics can be very close to each other.

A corpus of more than 1000 emails was taken for training and testing and a system. Pyam, has been implemented. Preliminary analysis shows a set comprising three principal speech acts can be recognized by this approach, although some more easily than others. One advantage of this general method is that it would be easy to adapt the approach by discovering patterns which occur more frequently in different domains. These patterns are found by counting the frequency of one, two, three-gram word patterns in the email domains. Previous work by Patel (1990) attempted to summarize email messages but, although she considered syntactic and semantic aspects of messages, she did not extract speech act content from them, modelling instead the effect of the contents on the reader. She analysed only a handful of messages, so her work remains unevaluated.

It is perhaps surprising, given the amount of work in Information Extraction of this general type (Jacobs 1992; Cowie and Lehnert 1996; ARPA 1996; Gaizauskas and Wilks 1997), and the pressing need to process email more effectively, that there is not more work in this area. Two large-scale searches of the Web and an appeal to a range of 50 bulletin boards suggest that there is nothing publicly available with which to make more detailed comparisons.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

3. Email

Email is a fast and easy tool for communication, and the language of email is, in general, an informal one. A receiver often deletes most messages as soon as he or she reads them. Most senders of email pay little attention to the way they type their messages, and bad spelling is normal in such messages. It is also a common habit to use non-standard abbreviations in their messages: using 'u' instead of 'you', 'pls' instead of 'please' and 'tnx' instead of 'thanks' along with simplified punctuation and smiley faces. All this makes it necessary to use some form of preprocessing of email before applying a more general, principled natural language processing system. There are two main types of pre-correction that Pyam applies to the messages:

(i) Abbreviations are replaced by complete words, as with 'please' for 'pls'.
(ii) Fillers like 'Ooops', 'um', 'uh' which usually occur in spoken dialogues and are found in email messages as well, and are deleted. These two types are recognized initially from hand-compiled lists included in the system.

3.1 Email text: spoken and written language

A corpus study by Halliday (1985) showed that spoken and written language have different lexical densities. Lexical density is a way to measure the frequency of lexical words in a text: if N is the number of words in a text and L is the number of lexical words in the text, then

$$\text{lexical density} = 100 \times \text{L}/\text{N}$$

Some reports on corpus analysis have shown that, on average, written texts have lexical density in range of 36 to 57, although a shopping list might have 100, and the lexical density in spoken texts is in average between 24 and 43.

To investigate the type (i.e. spoken vs. written) of email text based on its lexical density and compare the result with the lexical density of naturally occurring dialogue, dialogue; d93-9.1 from the TRAINS (Allen and Schubert 1993) corpus and a corpus containing 30 emails (again all sent to our departmental support group) were chosen. Both the dialogue and emails were chosen randomly, and the number of emails in this corpus was chosen such that the total number of words in the corpus and the dialogue were close enough to be comparable. All labels and text marks from the dialogue were removed before counting the total number of words. There are 356 lexical words (i.e. noun, adjective, adverb and main verb) in the dialogue, and 508 lexical words in the email corpus. The total number of words in the dialogue and in the corpus are 1136 and 1291 respectively. So:

Lexical density for the dialogue: $\text{LD}_\text{d} = 100 \times 356/1136 = 31$ per cent
Lexical density for the emails: $\text{LD}_\text{e} = 100 \times 508/1291 = 39.35$ per cent

Comparing LD_d and LD_g with the results reported earlier shows that the dialogue text is in the range of spoken texts, as expected, and email texts are closer to spoken than written language.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

4. Speech acts

The main idea behind speech act theory is that each utterance of a speaker is an action used to achieve certain goals. So, 'I promise to be home by midnight' could be uttered as part of a speaker's plan to get permission to go out. Speech acts were introduced by Austin (1962) and Searle (1969), and later defined in terms of REQUEST and INFORM by Cohen and Perrault (1979) as part of an AI program designed to analyse utterances in the domain of a railway station help system, and to generate answers helpful to passengers. The general speech act REQUEST, can be taken as covering the answering of questions and the making of requests to hearers (which are conventionally associated with the English syntactic forms of questions and commands, although it is widely agreed that the act can be performed under a wide range of surface forms).

We have taken a fairly conventional route and subdivided REQUESTS into three subtypes:

(i) Request-Action, which covers any request in which the sender tries to get the receiver to take an action as in 'Please close the door';
(ii) Request-Information, which covers requests in which a sender is looking for some sort of information, as in 'What is the time?';
(iii) Request-Permission, which has sometimes been seen as a subtype of either Request-action or Request-information, and covers requests in which a sender asks for permission to have some information or to perform an action, as in 'Can I put files on the group disk area?'

The differences between Request-permission and the other two is that in Request-permission the sender is not sure whether or not it is allowed to ask for that information or action. For instance, 'Please setup an MSc alias for me.' is a Request-action, 'How do I login to the department over the ethernet?' is a Request-information and 'May I have access to nlp directory?' is a Request-permission.

4.1 Indirect speech acts

Indirect speech acts are speech acts in which one such act is performed indirectly by way of performing another (Searle 1975). So if someone says 'Could you move over a bit please?', he does not expect a yes or a no reply, although it appears to be a yes/no question in syntactic terms. There are two types of indirect speech acts: the first consists of those indirect speech acts that are used frequently and whose literal meaning is normally not intended by the speaker. For example, although 'Can you please be quiet?' is literally a yes/no question, in almost all contexts it is an indirect Request-action, namely to BE quiet.

The second type of indirect speech acts is those utterances whose interpretations are heavily dependent on the context. For example 'Can you reach the shore?' is a yes/no question, and one would expect it to be a Request-information; yet the very similar 'Can you reach the milk?' would be the same in some contexts, in other contexts it is really an indirect Request-action.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

Searle (1975) defined six types of utterances which could quite standardly be used to make indirect Requests:

1. Sentences concerning the hearer's ability to perform actions.
2. Sentences concerning the speaker's wish or want that hearer will do action.
3. Sentences concerning the hearer's doing an action
4. Sentences concerning the hearer's desire or willingness to do the action.
5. Sentences concerning reasons for doing an action.
6. Sentences embedding one of the elements above inside another.

Although these groups categorize utterances which could be used to make indirect Requests, it is necessary to find out from data how users in different domains in fact express their requests. Our initial method is to discover patterns and phrases used in a test domain by counting the frequency of 1–3 gram words in the domain. This method, whose efficacy we believe our results confirm, can be contrasted with, on the one hand, methods like Hinkelman's (1989) that rest entirely on intuitively chosen cues to speech acts and, on the other, more recent work that seeks to derive cue combinations entirely from random combinations of features supplied to a machine learning algorithm. The best of such work (Samuel, Carberry and Vijay-Shanker 1998), has only achieved a 70% success rate in utterance tagging, and we believe that our methodology is reasonable in the domain we have chosen.

In addition, as Samuel *et al.* (1998) reported, one of the limitations of their strategy is time: '…the memory and time costs are so exorbitant that the method becomes intractable.' At present, such a method is not practical for any reasonable feature set.

5. Corpus analysis

For the purpose of this research, a corpus of 1000 email messages was prepared from a set of messages sent to the Computer Science Support Group at Sheffield University, Computer Science Department. Because of the nature of this domain, more than 90% of these messages are requests in their three different forms distinguished above. The first step is to assign speech act tags manually to the sentences of the corpus. The frequencies of significant 1–3 gram phrases occurring in sentences with particular speech acts are then computed. Tables 1 and 2 show the frequencies of the most frequent 1–3 gram phrases occurring as parts of Request-information and Request-action sentences of the corpus.

These frequency figures have been combined with other criteria gleaned from the 'speech-act-cue' literature to produce sets of criteria and rules for each speech act type under study. We shall now describe these in more detail.

5.1 Request-Information

Request-information occurs both directly and indirectly in this domain. The term 'inform-verbs' refers to the following verbs: 'tell', 'inform', 'reply' (in this domain) and 'advise'.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

Table 1. *The frequency of 1–3 grams in 300 Request-information messages*

1-gram		2-gram		3-gram	
phrases	freq	phrases	freq	phrases	freq
have	91	tell me	41	you tell me	30
can	90	is there	30	is it possible	15
go	90	could you	28	how do I	13
is	58	can you	24	do we have	12
?	55	do I	23	could you tell	11
tell	44	how to	18	tell me what	10
please	38	is it	18	tell me how	8

Table 2. *The frequency of 1–3 grams in 600 Request-action messages*

1-gram		2-gram		3-gram	
phrases	freq	phrases	freq	phrases	freq
please	400	could you	180	could you please	60
you	370	please could	68	please could you	53
could	325	can you	68	have a look	25
can	182	can I	18	please can you	12
have	139	please can	17	could someone please	10
me	90	can someone	16	would you please	10
will	60	would you	15	can you please	9

5.1.1 Direct Request-information

There are five types of sentences which are considered direct Request-information:

(i) wh-questions.
(ii) yes/no questions without modal verbs, e.g. 'Is there a reasonably fast 486 machine connected to the research network?'
(iii) yes/no questions with modal verbs and without 'please', e.g. 'can you include me to forum alias?'
(iv) Non-question form sentences which include 'please' and one of the inform-verbs followed by 'me' or 'us'. For instance 'please tell me how to print postscript files.'
(v) Non-question form sentences which include phrases like 'let me know', 'how can I', 'I want to know', 'any idea'. For instance 'any idea how I cope with this problem?'

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

5.1.2 Indirect Request-information

Literally yes/no questions with modal verbs plus 'please' plus any inform-verb plus pronouns 'you' as subject and 'me' or 'us' as objects is considered as indirect Request-information. In general, these sentences have the form: [please]{modal verb}you {inform-verb}{me,us} [please]. So while 'can you **put** my name on the list, please?' is indirect Request-action, 'Can you **tell** me about this list please?' is indirect Request-information and they only differ by their verbs.

5.2 Request-Action

Request-Actions occur both directly and indirectly in this domain.

5.2.1 Direct Request-action

Most direct Request-actions are identified by non-question sentences with the clue word 'please' plus a verb which is not an inform-verb, as defined above. For example 'Please restore my files in my directory'. Also requests with inform-verbs which are not followed by pronouns 'me' or 'us' are considered as Request-action. For example 'Please tell everybody about this seminar' is a Request-action and not a Request-information.

5.2.2 Indirect Request-action

Again, apparently yes/no questions with modal verbs and pronoun 'you' plus 'please' without inform-verbs, are indirect Request-action. For example, 'Can you please put my name in the c-users alias?'. Another class of indirect Request-action are yes/no questions with modal verbs plus 'please' plus inform-verbs followed by any noun or pronouns excluding 'me' or 'us'. For example, 'Can you please tell all students to be there?'

In general, inform-verbs can appear both in Request-action and Request-information. Consider these two requests: 'Please tell me your name' and 'Please tell him your name'. The first is a direct Request-information and the second a direct Request-action. One reason for this claim is that the first one can be substituted with a wh-question like 'What is your name?' which serves the same purpose, but there is no wh-question for the second one. To deal with this problem, the system should consider the nouns or pronouns that follow the verbs. The pronouns 'me' and 'us' when coming after inform-verbs indicate that the act is a Request-information.

5.3 Request-Permission

Although this speech act has some similarity with Request-information and Request-action, it constitutes a different speech act. For example, 'May I have access to /share/nlp/directly?' could be an indirect Request-action or indirect Request-permission. This ambiguity is largely one of UK/US dialect difference, since in the US, 'may' is reserved for request-action, e.g. 'Can/May I have the salt?' is a UK/US

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

```
The monitor on my machine is flashing like a faulty TV
screen. can some one look at it please?
<DOC>
<DOCNO>1234</DOCNO>
<p>
<s><COREF ID='1'>The monitor</COREF>on my machine
is flashing like a faulty TV screen.</s>
<s>can some one look at <COREF ID='2' TYPE='IDENT'
REF='1'>it</COREF> please?. </s>
</p>
</DOC>
The monitor on my machine is flashing like a faulty TV
screen.
can some one look at [The monitor] please?
```

Fig. 1. Use of LaSIE to substitute co-references.

alternation for the Request-action, and one on which Americans correct British speakers by saying 'Do you mean "may you have it?"' The 'may' form is largely Request-permission in the UK. Pyam therefore considers this request as indirect request-permission. Most Request-permission messages appear as indirect speech acts in this domain.

Direct Request-permission speech acts are recognized as non-question sentences, including 'please' plus verb 'let', plus pronouns 'me' or 'us'. Indirect Request-permission are yes/no questions starting with 'may' and followed by 'I' or 'we', if the rank code (see section 6.2) of the sender is lower than the rank code of the user. For example, 'please let me use your terminal' is a direct Request-permission and 'May I use your terminal?' is an indirect Request-permissions.

Another interesting point is the meaning of 'Indirect speech acts type two' (see section 4.1). This type of indirect speech act is very context-dependent, and is difficult to identify with a phrase approach. However in the domain analysed, some of these indirect speech acts occurred frequently enough to make it possible to identify them as indirect Request-actions. For example, 'rlab1 printer is out of toner' can be considered as a Request-action as long as we remember that our domain is emails to a support group, and the same message to somebody else can hardly be considered as a request.

5.4 Inform

In the analysis of the corpus, most of the inform speech acts act as either a pre-explanation or post-explanation of requests. 'Will you have a look at it? There appears to be a fault on the CD drive', or 'The printer queue of rlab2 seems to be stuck. Could you have a look at it?' are examples where (in the former) the inform

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

part of it serves as a post-explanation, and (in the latter) as a pre-explanation. Although these parts of messages do not generally contain the gist of the message, they can be used to substitute for any possible co-references in the request act.

Pyam uses the co-reference module of the LaSIE Information Extraction system (Cunningham, Wilks and Gaizauskas, 1996) which is part of a larger Information Extraction system. So in the first example, the system is able to substitute 'CD drive' for 'it', and in the second example it substitutes 'the printer' for 'it'. Figure 1 shows an original message, LaSIE's output and the substituted result.

6. Implementation

The implemented system Pyam, which is written in PERL, consists of four sections: email header lines, preprocessing, focus sentences and report generation.

6.1 Email header lines

Pyam picks up the three most informative items from header lines: Sender, receiver and subject (and ignores the rest).

'Receiver' is used to find out whether the email is a personal email or to a group. 'Sender' is checked against prepared files, based on the rank of the sender compared to the receiver, to help disambiguate issues of Request-permission; e.g. an undergraduate would normally be requesting permission if the receiver were an academic, but an academic would normally be requesting action if the receiver were a member of a support group. These ranking files are established in accordance with the individual rankings assigned by the end user of the system.

6.2 Preprocessing

Abbreviated words are replaced by their full forms and the sentences boundaries of the message are inserted. Sentence recognition, in this sense, is one of the most difficult problems, especially in this domain. Sentence separators are not used in standard ways in email, e.g. some standard separators like full-stops are used for different purposes. Other uses of the full stop are found in numbers (3.5), titles (Mr.), etc. Some people use uncommon separators like '-' and some use repeated separators, like '?????', and some do not use separators at all. Pyam splits a message into sentences by the sentence separators after recognition of titles, numbers and replacing multiple separators by a single ones.

The rank code of the sender compared to the receiver is determined in this stage. The three possibilities are higher, lower or unknown. This code is used to differentiate Request-permission from Request-action or Request-information.

6.3 Focus sentences

Sentences that match any of the patterns of section 5 are considered as focus sentences, in that they carry the nuclear message associated with the speech act

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

detected. Pyam picks up all the focus sentences of email messages by applying the criteria of section 5.

6.4 Report generation

Pyam assigns a speech act to each of the focus sentences of the message recognised as described above. Since parts of patterns overlap in different speech acts, Pyam has a hierarchy order ranking the most specific part of pattern first to the most general part at the end. For instance in the sentence 'Can you tell me about this list please?', considering 'please' as a common pattern between all request speech acts at the first match will not be helpful. Instead, Pyam first recognizes it as a yes/no question with modal 'can'; then its verb, 'tell', which is one of the 'inform verbs'; then its direct object which is 'me' and finally 'please'. Based on all the above information, Pyam considers the sentence to be an Indirect Request Information.

Pyam prints out partial messages containing the speech act of each focus sentence(s) as follows. The output for each email message has three parts: Header-lines, Focus sentences and speech acts results. The Header-lines normally contain Sender, Receiver and Subject. For each focus sentence, Pyam prints out its corresponding speech act type. So far, there are ten different speech act message types:

1. REQUEST-INFORMATION.
2. YES/NO QUESTION, REQUEST-INFORMATION.
3. WH-QUESTION, REQUEST-INFORMATION.
4. LITERALLY YES/NO QUESTION.
   INDIRECT REQUEST-INFORMATION.
5. REQUEST-ACTION.
6. LITERALLY YES/NO QUESTION. INDIRECT REQUEST-ACTION.
7. REQUEST-PERMISSION.
8. LITERALLY YES/NO QUESTION.
   INDIRECT REQUEST-PERMISSION.
9. REQUEST PERMISSION or REQUEST ACTION.
10. REQUEST PERMISSION or REQUEST INFORMATION.

These messages express more information than just the more general speech acts types at the coarser grain of the three REQUEST sub-types. This information functions now as a diagnostic tool for improving the systems detection mechanisms. It is not part of what would be given to a potential user, who presumably has no interest in mappings of syntactic to speech act structure; nor does it function as part of the evaluation mechanism, since the original tagging is, of course, to the content or speech act type, and not to its indirection type or underlying syntax. Messages 9 and 10 are reserved for situations when the rank of the sender is unknown and the final decision is left to the receiver to make.

There are two other messages output by the system. The first one is: 'This message has no REQUEST speech act.' which appears when Pyam finds no request, and the second one is 'NOT KNOWN', which appears when a conventional phrases that

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

allows it to be categorized as a Request. In a message like, 'I used to print file from Word for Windows in the PC before but now I can't!!?', although the sentence is really a Request-information, Pyam is unable to recognize the speech act. Figure 2 shows an original email message and the results from Pyam.

```
From Ted
Date: Mon, 16 May 94 16:12:34 BST
From: Ted
To: support
Subject: question re.solaris 2
Content-Length: 109
Status: RO

Please tell me what issues would be involved in
installing solaris 2.4 on a machines. there's some sun
software that doesn't really get supported for s1
anymore. Could you please send me any related
documents?
  SENDER: Ted
  RECEIVER: support
  SUPPORT: question re.solaris 2

1-Please tell me what issues would be involved in
installing solaris 2.4 on a machines.
    REQUEST-INFORMATION.

2-Could you please send me any related documents?
    LITERALLY YES/NO QUESTION.
    INDIRECT REQUEST-ACTION.
```

Fig. 2. Original email and the result prepared by Pyam.

7. Evaluation

Pyam has been evaluated in three phases.

7.1 Emails from the same domain

A new corpus was prepared from email to the Computer Support Group, but different from the corpus that was used to formulate the criteria of section 6. The focus sentences of 24 messages were underlined, and the speech acts for each of the focus sentences were assigned by three volunteers.

In speech act recognition, all three volunteers agreed with Pyam about 19 of the 24 speech acts in the focus sentences. In three of the messages two of them agreed, and in two speech acts only one of the volunteers agreed. Table 3 shows that 90.28 % of the results captured by Pyam are correct.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

Table 3. *Pyam's first evaluation result*

No of volunteers agree in with Pyam	No. of messages for which they agreed	Grades	Per cent
3	19	57	79.17
2	3	6	8.335
1	2	2	2.78
Total	24	65	90.28

Table 4. *Pyam's second evaluation result*

	Manually	PYAM				Results
	Recognized	Recognized	Correct	Wrong	Missed	Per cent
F.S.	116	116	108	8	8	95.73
S.A.	116	91	82	9	21	84.58

7.2 Emails from different domains

The second phase of the evaluation differs from the previous one in the nature of the message receiver. The system was originally trained on emails to the Technical Support group and tested by different emails to them. For this evaluation we took emails to a member of the staff in the Department of Computer Science. There were 100 random emails to him which varied greatly in length and subject.

Table 4 shows the results prepared manually and by Pyam over these emails. The first row shows the results for Focus Sentences (FS), and the second row shows the results for Speech Acts (SA).

From the 100 messages, the Requests in 79 of them have been recognized correctly both for FS and SA, which demonstrates reasonable generality across email domains by Pyam.

7.3 Spoken dialogues

To evaluate the efficiency of Pyam in different domains further, another corpus has been tested: TRAINS (Allen and Schubert 1993), which was collected as part of the TRAINS project (Allen, Ferguson, Miller and Ringger 1995).

This corpus contains task-oriented spoken dialogues: 98 dialogues, about 5900 speaker turns, and 55,000 transcribed words. The main problem dealing with the TRAINS corpus is the way that dialogues have been segmented: utterances can have more than one sentence without any end of sentence indicator. Pyam was able to recognize all requests which appeared as defined in the patterns described in the previous sections, but unable to capture requests embedded in another sentence. For example, in dialogue d93-26.3

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

utt1: s: hello can I help you

utt2: u: yes um <sil> to take two boxcars <sil> with the <sil> two engines from Elmira <sil> to Corning <sil> is <sil> would be how many <sil> hours

Pyam is able to recognize "*can I help you*" as a request but is unable to recognize "*is would be how many hours*" as a request-information.

The requests recognized by Pyam in this domain are all in direct Request-information form. Unrecognizable questions are those either embedded in a sentence, e.g. utt23:

utt23: u: um hm next uh <sil> hm <sil> okay I'd like to send <sil> uh does it take any less time if it's just an engine <laughter>

or are in question form because of their intonations, e.g. utt11:

utt11: u: okay uh <sil> only one train can be on the track at one time

There are 21 requests for information in this dialogue, of which Pyam is able to recognize 15, which is more than 71%.

These figures compare favourably with the levels of results reported by Samuels *et al.* (1998) using machine learning methods for speech act tagging.

8. Conclusion

Although phrase matching is not a strongly theoretical approach for identifying speech acts, it still performs well against other methods and, with some additional pragmatics and linguistic components, it should be able to locate the gist of messages, especially in a pre-defined domain. The results captured by Pyam show that most Requests (information, action and permission), expressed both directly and indirectly, can be recognized by this approach. It would be relatively straightforward to extend this approach to capture the differences between conference calls for papers, announcements, results of paper submission, and so on, so as to provide a service to users that captured more content, as well as the broad routing of speech act types we offer here. Standard information extraction modules for dates and places could augment this fairly straightforwardly.

Acknowledgements

One of the authors (H. Khosravi) expresses his thanks to the MCHE of the Islamic Republic of Iran for granting a scholarship for a PhD in computers, during which this work was carried out. We are grateful for helpful suggestions and comments from K. Bontcheva, N. Dahlan and M. Lee, and for their help in evaluating the system; the mistakes are, of course, our own.

References

Allen, J., Ferguson, G., Miller, B. and Ringger, E. (1995) Spoken dialogue and interactive planning. *Proc. of the ARPA Spoken Language Technology Workshop*, Austin, TX.

Allen, J. and Schubert, L. (1993) Language and discourse in the TRAINS project. In: Ortony, A., Slack, J. and Stock, O. (editors), *Communication from an Artificial Intelligence Perspective*. Heidelberg: Springer-Verlag.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press

ARPA (1996) *The Sixth Message Understanding Conference* (*MUC-6*), San Francisco, CA. Advanced Research Projects Agency (ARPA), Los Altos, CA: Morgan Kaufmann.

Austin, J. (1962) *How to Do Things with Words*. Oxford: Oxford University Press.

Cohen, P. and Perrault, C. R. (1979) Element of a plan-based theory of speech acts. *Cognitive Science* **3**(3): 177–212.

Cohen, W. (1995) Learning to classify English text with ILP methods. In: Raedt, L. D. (editor), *Advances in Inductive Logic Programming*. Amsterdam: IOS Press.

Cowie, J. and Lehnert, W. (1996) Information extraction. *CACM* **39**(1): 80–91.

Cunningham, H., Wilks, Y. and Gaizauskas, R. (1996) GATE: an environment to support research and development in Natural Language Engineering. *Proc. COLING 96*, Copenhagen.

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

Guthrie, L., Walker, E. and Guthrie, J. (1994) Document classification by machine: Theory and practice. *Proc. COLING 94: The 15th International Conference on Computational Linguistics* (2): 1059–1063.

Halliday, M. (1989) *Spoken and Written Language*, *Second Edition*. Oxford: Oxford University Press.

Hinkelman, E. (1989) Linguistic and pragmatic constraints on utterance interpretation. *PhD thesis*, Computer Science Department, University of Rochester, NY.

Jacobs, P. (editor) (1992) *Text-based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. Hillsdale, NJ: Lawrence Erlbaum.

Lee, M. and Wilks, Y. (1996) An ascription-based approach to speech acts. *Proc. COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen.

Patel, M. (1990) The Summarisation and Categorisation of Electronic Mail Messages. *PhD thesis*, University of Sheffield.

Samuel, K., Carberry, S. and Vijay-Shanker, K. (1998) Dialogue act tagging with transformation based learning. *Proc. COLING-ACL 98: The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal.

Searle, J. (1969) *Speech Acts*. Cambridge: Cambridge University Press.

Searle, J. (1975) Indirect speech acts. In: P. Cole and J. Morgan (editors), *Syntax and Semantics*, (*3*): *Speech Acts*. New York, NY: Academic Press.

https://doi.org/10.1017/S1351324999002119 Published online by Cambridge University Press