Determining query types for information access

Simon Smith and Martin Russell
School of Electronic and Electrical Engineering, University of Birmingham
smithsgj@eee.bham.ac.uk

Abstract

A body of research exists on the statistical characterization of speech acts, drawing on the features of the utterance at various linguistic strata. If queries to a multi-functional information access system could be similarly analysed, it might be possible to determine, without user intervention, the module to which control should be passed for a given task. Query determination is taken to be a special case of the general multi-class problem, also exemplified by topic spotting. Therefore, after introducing some previous work on the problem and providing motivation for the present research, the paper describes a preliminary topic spotting experiment, along with the results. Future directions for the query typing work are then set out.

1. Introduction

Most of those involved in information research are actively engaged in designing Information Access tools. Information retrieval, information extraction, text mining, data mining, spoken document retrieval, named entity extraction: all are concerned, in the end, with the mapping of some user query on to some set of relevant data, and passing that data back to the user.

What we wish to undertake is probably less ambitious. We assume that there exists a comprehensive information access system: discrete components of the system can retrieve documents, extract information from documents, retrieve data from fixed format files (for example the on-board address book or calendar manager), and perform certain user requested actions, such as connecting a phone call. The system would probably be mounted on a PDA-type device, and incorporate a voice input facility. We further assume that the system, for all its capabilities, is entirely piecemeal: what is lacking is a central control or shell mechanism which can determine, from the form of the user query, which of the four components is right for the task at hand.

What we shall be attempting to do, therefore, is to examine the linguistic evidence present in a query, and on the basis of that evidence determine the query *type*. This task is similar in many respects to the assignment of *speech act type*, *utterance type*, *illocutionary force*, or *dialogue act type* to an utterance in Discourse Analysis; an important strand of the work, therefore, will be to investigate the applicability of traditional linguistics insights to a very modern problem.

Research on speech act classification, as well as the strategy proposed for query typing, is closely paralleled by work on *language identification* (LID). If it is assumed that word boundaries are known, as is the case for text LID, “the algorithm used can be derived from basic statistical principles [and] is not based on hand-coded linguistic knowledge, but can learn from training data”, according to Dunning (1994). Interestingly, Dunning also notes that the same algorithm has been implemented for *species identification* in the domain of biochemistry; moreover, his keyword-counting approach bears comparison to Garner’s (1997) dialogue act classification.

Such techniques can be generalized to a means of solving other multi-class problems: Gorin (1995) worked on a method of determining the topic of a telephone call, so that it could automatically be routed correctly. This is an appropriate application for LID, too: for example, in a multilingual culture, it would be helpful if calls could be directed to an operator fluent in the language of the caller (Lazzari, Frederking and Minker 1999). Now, imagine a machine translation system which consists of components for various language pairs. If the source language were not specified by the user, it might be determined by a keyword-based language spotter, so that control could be passed to the requisite MT component.

In both these cases, then, we have something akin to a topic spotter acting as a command or shell function, determining the correct module for the task itself. The same idea can be extended to information access: a personal digital assistant might have a limited number of command and query strategies at its disposal, each employing a distinct technology. The choice of strategy would, in principle, be governed by the user’s expectation of the results of the query or command, so one approach would be to require him or her to select a query type before entering the query itself. Far better, though, and more in accordance with one’s expectations of a personalisable device, if the right module could be determined from features of the query itself.

As well as addressing the practical matter of an approach to a real world problem, the work is also an exercise in statistical language analysis. We plan to consolidate the extension of topic-spotting to multi-class problems in general, and to the determination of query types in particular; then build from training data a set of convenient query features at various significant linguistic strata.

2. A query typology

What, then, are the possible query types, and what applications may they be associated with? Possible types might include *command action*, *open exploration*, *factual enquiry*, *how-to-do enquiry*, *buying enquiry*. Many of these, it turns out, can actually be decomposed into what might be termed *query primitives*: a buying enquiry, for example, could be taken to consist of open exploration, factual enquiry and command action phases.

Open exploration corresponds to what is technically known as *information retrieval*. As with a web search, it is not answers to questions that are returned, but details of (or hypertext links to) documents in which the user might find the answers. IR retrieves pointers to information, not the information itself.

Factual enquiry seeks dates, names or other data in direct response to the query. Two subtypes can be distinguished: the first exploits the relatively new technologies *question answering* and *information extraction*, where the contents of texts and other documents are analysed and data returned. Far more straightforward is *data retrieval*, or database enquiry: this would apply to local files of known formats, such as the address book or calendar manager found on all PDA and similar systems.

Command action needs no elaboration, beyond stating that it is not strictly a “query” type, and the task application is one of synthesis rather than analysis.

Example queries of each type are now presented.

(1) IR we need a map of/information on Manchester Airport.

(2) IE What terminal does Singapore Airlines use at Manchester Airport?

(3) DR What’s the name of the guy I’m meeting at the airport on Tuesday?

(4) CA Call him, please.

It is immediately clear that there is little in the way of an intuitive connection between the query type and the vocabulary used. It is true, perhaps, that one might associate *information* with the IR type, wh- question words with factual enquiries, and *call* or *please* with command actions; but it is easy to think of counter-examples, and most real queries will not contain any of these keywords. In any case, the examples are not real queries but have been invented by us.

The goal, however, is not to handcraft rules of association, but to observe whether, over a reasonably large training set, any patterns do emerge; whether, and to what extent the query type can be predicted on the basis of query keywords.

3. Lexical speech act determination

Samuel et al (1998) used a *transformation-based learning* (TBL) heuristic algorithm to assign speech act type to texts. They derived a set of ordered rules and applied them in turn, so that each utterance was first labelled with a default speech act tag. The application of subsequent rules led to utterances containing known keywords (*dialogue act cues*) being re-tagged. In their data, over 71% of utterances were correctly tagged by this means. Samuel et al noted limitations in their TBL model, namely that the rule-templates had to be hand-crafted, and that the algorithm could suggest speech act tags but had no means of asserting its confidence in those tags; they identify a possible solution for the first problem and a workaround for the second.

Lager and Zinovjeva (1999) applied the TBL algorithm to the Edinburgh Map Task, a corpus which incorporates dialogue act mark-up, and attained an accuracy of 62.1%. The corpus consists of a series of conversations about the contents and features of a map, and each utterance is marked up for dialogue move: *Go to the left* is an INSTRUCT move, *What’s behind the church?* is classified as QUERY-W, *Right* and *OK* as ACKNOWLEDGE, and so on. In all, there are 12 different dialogue moves.

Garner (1997) undertook the same task, based on the same corpus, using a maximum likelihood model. Garner reports a classification success rate of 47%, which does not sound particularly promising until it is recognized that his model treats both moves and keywords as independent, that is to say no account is taken of contextual information. Furthermore, the approach is based purely on lexical frequency, and no appeal is made to other linguistic strata.

The annotation of the Map Task corpus is reported on by Carletta et al (1998).

There is a clear parallel in Garner's experiments with the query typing work here proposed. Furthermore, since we propose only four distinct query types, it is reasonable to expect a higher classification accuracy.

4. Experiment – a topic spotter

As an initial experiment in multi-class problem solving, we built a baseline *topic-spotting* system, implementing algorithms of (and effectively replicating the work of) Garner (1997) and Wright et al (1995). The approach is rather different from that of standard IR systems or web search engines: it attempts to find the single best solution to a multi-class problem from a very constrained set of candidates, or alternatively to rank each member of that set. IR, on the other hand, offers in general a ranked list of possible solutions from a virtually unlimited set of possibilities. In principle, however, the topic-spotter could be extended to handle a very much larger topic set.

Our aim was to build a system which would correctly identify the topic of an unseen news text, by examining the frequencies of significant keywords used. To do this, we first decided on ten appropriate topics: domestic politics, foreign politics, health, finance, crime, sport, media, technology, environment and education. Two sets of training data were assigned to each category, one set consisting of eight news stories, the other — a subset of the first — of three stories. All the texts were taken from BBC Online News, where they are arranged in categories not dissimilar to ours; the fact that multiple category membership of texts is commonplace in the BBC scheme is what led us to develop our own. Often, though, the topic of a text seemed to motivate its assignment to more than one class (a story about Elizabeth Taylor and her AIDS-related work belonged intuitively to both media and health categories, for example). In such cases, we simply excluded the text from training data.

Once training data had been prepared, the probable topic of an unseen text was computed by maximizing probability (5) (Garner 1997) over all keywords and all topics:

$$(5) \quad P(\mathrm{x}|m_i,\mathrm{D}) = P(w_1|m_i,\mathrm{D})P(w_2|m_i,\mathrm{D}) \ldots P(w_K|m_i,\mathrm{D})$$

Here, $P(\mathrm{x}|m_i,\mathrm{D})$ represents the probability that an observation *x*, instantiated as the unseen text, will occur in the training data D associated with a topic *m.* Whenever a word encountered in the unseen text is found in the training data for a particular topic, its probability of occurrence ($P(w_j|m_i,\mathrm{D})$) is computed by dividing the number of tokens of that word in D by the total number of words in D; if D does not attest the word in question, the probability defaults to an arbitrarily small value. As (2) shows, these probabilities are then multiplied together to yield $P(\mathrm{x}|m_i,\mathrm{D})$. Normally, $P(\mathrm{x}|m_i,\mathrm{D})$ would in turn be multiplied by the topic prior probability before maximization, but that step is skipped in this implementation, as it is assumed that all topics are equally likely.

When the small (three stories per topic) training set was used, 30 out of 50 stories were assigned to the right topic, while 42 stories were correctly matched by means of the eight-story training set.

4.1 Data pruning

One of the basic principles on which this work is founded is the use of information-theoretic measures, such as *usefulness* (Garner 1997) or *salience* (Gorin 1995) to identify structure which is, in some sense, optimal for discriminating between the different classes in question. These techniques are most commonly applied at the word level, either to text, verbatim transcriptions of speech, or the output of a speech recognition system. However, in principle they are equally applicable to any symbolic representation of data. For example Gorin et al (1999) have applied similar techniques to the output of an automatic phone recognizer. Of course, first-order statistics describing the occurrence of individual phones are unlikely to provide strong evidence for the classification of an utterance. However, by extending these techniques to phone sequences, it may be possible to detect, automatically, phone sequences which describe discriminative lexical or even syntactic or semantic structure. Potentially useful sequences will be characterized by their frequent occurrence in different contexts and can be detected automatically using methods such as the Context-Adaptive Phone (CAP) analysis described in Moore et al (1994). Variations in the instantiation of the same underlying sequence, due to deletion, insertion or substitution of symbols can be accommodated using various schemes based on dynamic programming (Sankoff and Kruskal 1983), and this could be formalized, for example, by representing each sequence as a statistical model, such as a hidden Markov model (Jelinek 1999). Similar techniques could be applied to sequences of primitive acoustic symbols, produced by a data-driven cluster analysis of the output of a speech signal processing system, or, as in the case of the research proposed here, to more symbolic representations such as those used to describe prosodic information, such as that described by Shriberg et al (1998).

Whilst the experiments conducted so far have been confined to lexical analysis, the ultimate goal is to process input from an unsegmented speech stream. Thus, it is intended to build models based on sub-word units, probably phones.

4.1.1 Usefulness thresholding

Our program then applied (6) (Wright et al 1995) to determine a *usefulness* score for each vocabulary item in the training data.

$$U_k = P(w_k|\mathrm{T})\log\frac{P(w_k|\mathrm{T})}{P(w_k|\bar{\mathrm{T}})} \qquad (6)$$

A word *w* is thereby said to be useful when it is frequent in training texts of topic T, and occurs relatively rarely in other texts; the usefulness score describes the discriminatory contribution of keywords to the topic of the text. Thresholding or pruning is then carried out so that only the *n* most useful keywords are searched for when determining the topic of an unseen text.

Table 1 shows, for each news topic, what were computed by the usefulness algorithm to be the top ten keywords. The lists probably correspond to most people's intuitions of words that would epitomize each topic.

Table 1

finance	*computers*	*crime*	*domestic politics*	*education*	*environ ment*	*foreign politics*	*health*	*media*	*sport*
banks	internet	plane	party	students	environ ment	eritrea	cancer	film	boxing
yen	web	victim	tories	schools	masts	lazio	tamoxifen	olsen	sydney
merger	websites	suharto	kennedy	college	gm	ethiopia	breast	laurel	olympic
bank	engines	bomb	labour	oxford	fuel	eritrean	parodi	travolta	ham
jp	lottery	victims	mp	university	crops	speight	fruit	films	spurs
shares	computer	cheng	ira	pupils	we	monitors	everson	magazine	garcia
sega	information	police	donaldson	curriculum	oil	kosovo	boots	hardy	talent
debenhams	sites	trial	ulster	comprehensive	pioneer	fiji	krishnamurthy	hurley	robson
chase	neurons	musharraf	trimble	state	environ mental	ethiopian	removed	comedy	mcgrath
banking	identity	li	romsey	i	prince	electoral	pacemaker	book	edwards

However, when topics were assigned to unseen (test) news stories, greater accuracy was on the whole achieved when all training data was considered than when thresholding was applied. Table 2 shows how many of 50 stories were correctly identified when usefulness thresholding was applied at various levels *n*: that is to say, when only the *n* most useful words in the training corpus for each topic were taken into account.

Table 2

threshold	*usefulness*	*modified usefulness*	*salience*
10	10	30	10
20	13	31	10
30	12	35	15
40	12	35	19
50	13	37	18
60	12	39	24
70	11	38	28
80	12	35	30
90	13	34	30
100	13	34	29
150	12	35	31
200	16	36	34
249	16	37	35
299	16	37	36
349	16	38	36
399	16	42	42
449	16	39	42
499	19	41	43
549	22	41	42
599	23	41	40
no threshold	42	42	42

Table 2 also shows the results of thresholding under a reformulation of the usefulness calculation which we term *modified usefulness*, also suggested by Garner (1997). There is a marked improvement here, although the best performance is still achieved when no thresholding is carried out. The reformulation, shown at (7), differs only from (6) in that the absolute probability of occurrence of a word is ignored.

$$(7)\quad U_k = \log \frac{P(w_k|\mathrm{T})}{P(w_k|\bar{\mathrm{T}})}$$

The *salience* results will be discussed presently.

Part of the discrepancy between the results from the two formulations of usefulness arises because the original form tends to rank function words unduly highly, which leads the application to treat these items as keywords. Table 3 shows the ranking of the determiner *the* for each training corpus, under both formulations of usefulness.

Table 3

corpus	*modified*	*original*
computer	931	1032
crime	1026	1132
domestic	896	32
education	946	146
environment	1101	172
finance	912	1012
foreign	928	29
health	1038	518
media	951	1046
sport	1003	957

4.1.2 Salience thresholding

Gorin (1995) postulates *salience* as "an information-theoretic measure of how meaningful a word is for a particular device [i.e. task]". Gorin's work focused on unconstrained speech driven routing of telephone calls to one of a number of human operators, each dealing with one task, such as reverse-charge calls, credit card calls and directory enquiries. Salience of a word to a class *T* is computed by (8).

$$\text{(8)} \qquad \mathrm{sal}(w) = P(\mathrm{T}|\,w)\log\frac{P(\mathrm{T}|\,w)}{P(\mathrm{T})}$$

P(T) is ignored for our purposes, as all topics are equally likely. We do not know the probability of the class given the word, but it can be derived from Bayes' Law, as shown at (9).

$$\text{(9)} \qquad P(\mathrm{T}|\,w) = \frac{P(w|\mathrm{T})\,P(\mathrm{T})}{P(w)}$$

P(T) is again ignored; *P*(*w*) is the number of tokens of a particular word in the whole of the training data (in principle divided by the total number of tokens of any word, but this is a constant); *P*(*w*|T) is the number of tokens of the given word divided by the total word count for the topic, as per the usefulness calculation.

It will be seen that whereas usefulness compares the likelihood of a token in on-topic and off-topic texts, salience relates incidence in on-topic texts and all texts, whether on- or off-topic. As Table 3 shows, at low pruning thresholds, the algorithm performs as badly as the original usefulness formulation, but the performance accelerates towards that of the modified usefulness.

We are not unduly discouraged by the poor performance of the data pruning algorithms. They were implemented by Gorin and Garner with short utterances, rather than news stories, in mind; we will in due course be adapting these models to handle queries, which are by their nature short.

What we have attempted at this stage of the research is a baseline maximum-likelihood classifying program, and we are satisfied that it is working correctly.

5. Speech act determination from non-lexical features

We plan, next, to begin building other query features into the model. Of the knowledge sources described above, we expect discourse context, the likelihood of the next query type given the current selection, to prove the least difficult to integrate. This is because it involves the recycling of information that the program was designed to provide – the value assigned to the first utterance is simply passed as a parameter in the computation of the next, and, unlike the prosodic input, does not have to be determined externally.

5.1 Evidence from discourse context

Nagata and Morimoto (1993) trained a corpus of conference-booking dialogues using a trigram model of utterances classified by speech act type, and attempted to predict, using *mutual information*, the following utterance type given the current one. Their approach is entirely probabilistic, and ignores any linguistic or lexical considerations: they report a classification accuracy of 61%, amongst the nine speech act types catered for.

One would expect, certainly, patterns of speech act type to emerge at the discourse level: given a question, it is reasonable (if facile) to suppose that the next utterance will be an answer. An example of a sequence of query types was noted above with reference to the "buying enquiry", and (1) to (4) were intended to convey some sort of logical progression in an imaginary user's query trail. Whatever "answer" the query may generate is not part of the process here, of course, but there is no reason to suppose that monologue cannot be just as effectively modelled, statistically, as dialogue.

A couple of drawbacks would apply to this approach to query typing. First, in an isolated query, there is no "next" or "previous" utterance; and if there is, how does one decide whether contemporaneous queries are in fact related? And how recent does the previous query have to be for it to count?

Jurafsky et al (1997) labelled a portion of the Switchboard telephone corpus with 42 different utterance types. They combined discourse grammar (estimation of type based on adjacent utterances, as with Nagata and Morimoto) with keyword-based classification, and claimed accuracy of 64.6% (compared to 42.8% for their own keyword-only classification) in assigning utterance types to a test set.

6. Evidence from prosodic features

Jurafsky et al took into account information from another linguistic stratum in their classification: namely, prosodic features of the speech stream. They set up several dozen utterance-wide prosodic feature types, including *f0_max_utt* (the maximum fundamental frequency reached), *rel_nrg_diff* (ratio of RMS energy of final and penultimate phrasing region), and *mean_enr_utt* (mean speaking rate value); then they trained CART-style decision trees (Breiman et al 1984) whose task was to distinguish between two utterance types. The path through the tree, and eventual classification, was achieved through comparison of the feature values to constants stipulated at decision points. In the proposed work, we would initially adopt Jurafsky et al's prosodic features.

The CART tree-building technique employs *binary recursive partitioning*. Given training data (speech segments, the dialogue act types they represent and the prosodic feature values associated with them), configurations of parent nodes with two children are hypothesized. Each bifurcation represents a decision point on a particular feature; the algorithm examines all possible values for that feature, and attempts to find a *splitting rule* that maximizes the discriminatory power of the feature. Thus, while the choice of features is made by the experimenters, the trees are derived by data-intensive means.

Although the incorporation of prosodic information secured only a marginal improvement in performance, it is fairly well motivated, as there is a significant literature in the Discourse Analysis domain of theoretical linguistics on the relationship between intonation and speech act type (for example Crystal (1969), and Brazil (1985) for the Birmingham *discourse intonation* perspective).

The chief success of Jurafsky et al, in recruiting a prosodic contribution, seems to spring from the observation that, in English, a yes/no question tends to end with rising intonation. Queries to information access systems probably only rarely take the form of a yes/no question, so one might justly be sceptical about the application of prosodic techniques here: it is reiterated, however, that the aim of the work we propose is to establish what patterns of usage, if any, do emerge, with respect to the query types described.

There is a body of related work exploiting prosodic features in speech recognition, including that of King (1998), who integrates language model, dialogue context and intonational information to improve recognition of spontaneous dialogue speech. Jensen et al (1994) present a scheme for phrase-level recognition of intonation contours, and show how it can help compute the perceived pitch of voiceless utterance segments (where no fundamental frequency measurement is available). Hirschberg et al (1999) found that prosodic features can be used to predict recognition errors: their work was based on dialogues from the TOOT train information corpus. Carey et al (1996), working on speaker identification, established that prosodic features are less sensitive to noise distortion than cepstral coefficients.

7. Conclusions and future work

The next logical step is the extension of the topic spotter to determine dialogue moves in the Map Task corpus, following Garner, and this work is now under way. The extraction of annotated utterances to move-specific training corpora, using XML tools, is complete; some preliminary tests with held-out data indicate that usefulness has more of a role to play in dialogue move assignment than with the topic-spotter. Experiments using bigrams, rather than single words, will in due course be conducted.

The Map Task is not, however, a collection of queries, and identification of suitable query corpora is a high priority. The ATIS (Airline Travel Information Service) is a possible candidate, as are query logs from IR systems, some of which are publicly available on the web.

Once an appropriate corpus has been located, we will almost certainly have to do substantial annotation work. We plan to annotate using XML tags, for two reasons: first, we are currently working with the XML version of the Map Task, so fewer modifications to the dialogue move program will be needed. Secondly, the use of what is likely to become the standard annotation of choice would mean our work was more likely to be of use to other researchers, in the future.

It may be found more practical to use more than one corpus, although there could be methodological difficulties here, particularly if each corpus held only one query type (at the very least, the discourse context approach to query typing would thus be ruled out). Crucially, though, the corpora must represent true queries to either an on-line or Wizard of Oz system, rather than likely-sounding utterances coined for the purpose.

It may prove necessary to filter out very short queries: it might be difficult to distill any features at all from a query which consists of just one keyword, where the discourse context, prosodic and syntactic analyses are apparently unavailable. On the other hand, we might find that *query length* has a direct bearing on query type.

In the longer term, we shall be collecting data on the discourse context and prosody of spoken queries, to supplement the key word/phone information. Ultimately, we hope to employ data fusion techniques to develop a model of query type determination based on all these linguistic sources.

Once this has been completed to our satisfaction, we should like to consider exploiting a further source of information present in the query term which we have not discussed here in any detail. Intuitively, it seems plausible that the *syntactic structure* of a query may reveal information about its nature: a command action may be more commonly associated with the use of the imperative form of the verb, for instance, than other types, and one might expect a factual enquiry to invoke a more complex grammatical structure, perhaps including *long distance dependencies*.

Gorniak (1998) used a syntax based feature extraction algorithm to determine the topic of email messages. It is likely, too, that other topic-spotting techniques may make implicit use of syntactic information, perhaps through n-tuple based methods or statistical grammars. This possibility will be investigated further if it is decided to proceed with the syntactic work.

However, we know of no work so far that has attempted to evaluate or use the syntax of short texts or, specifically, queries. Such work would be very interesting and challenging. It is true that parsers and part of speech taggers are available and freely downloadable, and assertions about syntactic structure can be made with more confidence than is usual for *prosodic* structure, but the difficulty lies in establishing which features are to be taken into account, and what weighting is to be assigned to them.

The work we have embarked on is of potential importance to a number of disciplines. Information Access will benefit from the work, although it is not in itself an IA application; the prosody findings may be of general application in Speech Recognition; ultimately, it is envisaged, the work could form part of a production PDA device. It is, we believe, a novel and interesting approach, and can reasonably be expected to make a central contribution to the state of the art.

Acknowledgements

Thanks are due to Graham Tattersall of Snape Signals Research for stimulating discussion on queries and query typology; to Harvey Lloyd-Thomas of Ensigma Technologies for help with BBC Online News stories; and to Steve Isard of Edinburgh University for supplying the annotated MapTask, and for useful advice.

References

Breiman L, Friedman R, Olshen R, Stone C 1984 *Classification and regression trees.* Pacific Grove CA, Wadsworth.

Brazil, D 1985 *The communicative value of intonation in English.* Birmingham, University of Birmingham English Language Research.

Carey M, Parris E, Lloyd-Thomas H, Bennett S 1996 Robust prosodic features for speaker identification. In *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, pp 1800-1803.

Carletta J, Isard A, Isard S, Kowtko J, Doherty-Sneddon G, Anderson A 1998 The reliability of a dialogue structure coding scheme. http://www.cogsci.ed.ac.uk/~jeanc/maptask-coding-html/rerevised.html.

Crystal D 1969 *Prosodic systems and intonation in English.* Cambridge, Cambridge University Press.

Dunning T 1994 Statistical identification of language. Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University. http://www.comp.lancs.ac.uk/computing/users/paul/ucrel/papers/lingdet.ps.

Garner P 1997 On topic identification and dialogue move recognition. *Computer Speech and Language* 11(4): 275-306

Gorin A 1995 On automated language acquisition. *Journal of the Acoustical Society of America* 97: 3441-3461

Gorin A, Petrovska-Delacrétaz D, Riccardi G, Wright J 1999 Learning spoken language without transcriptions. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

Gorniak P 1998 Sorting email messages by topic, University of British Columbia Computer Science Dept. http://www.cs.ubc.ca/spider/pgorniak/um/bucfe.html

Hirschberg J, Litman D, Swerts M 1999 Prosodic cues to recognition errors. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

Jelinek F 1999 *Statistical methods for speech recognition.* Cambridge MA, MIT Press.

Jensen U, Moore R, Dalsgaard P, B Lindberg 1994 Modelling intonation contours at the phrase level using continuous density hidden Markov models. *Computer Speech and Language* 8(3): 247-260.

Jurafsky D, Bates R, Coccaro N, Martin R, Meteer M, Ries K, Shriberg E, Stolcke A, Taylor P, Van Ess-Dykema C 1997 Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, pp 88-95.

King S 1998 Using information above the word level for automatic speech recognition. Unpublished PhD thesis, Edinburgh University.

Lager T, Zinovjeva N 1999 Training a dialogue act tagger with the µ-TBL system. In *Proceedings of the Third Swedish Symposium on Multimodal Communication*, Linköping.

Lazzari G, Frederking R, Minker W 1999 Speaker-language identification and speech translation, *Multilingual information management: current levels and future abilities*, Pittsburgh, Carnegie Mellon University Computer Science Dept.

Moore R, Russell M, Nowell P, Downey S, Browning S 1994 A comparison of phoneme decision tree (PDT) and context adaptive phone (CAP) based approaches to vocabulary-independent speech recognition. In *Proceedings of the International Conference On Acoustics, Speech, And Signal Processing*, Adelaide, I: 541-545.

Nagata M, Morimoto T 1993 An experimental statistical dialogue model to predict the speech act type of the next utterance. In *Proceedings of the International Symposium on Spoken Dialogue*, Tokyo, pp 83-86.

Samuel K, Carberry S, Vijay-Shanker K 1998 Dialogue act tagging with transformation-based learning. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, pp 1150-1156.

Sankoff D, Kruskal J 1983 *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* London, Addison-Wesley

Shriberg E, Bates R, Taylor P, Stolcke A, Ries K, Jurafsky D, Coccaro N, Martin R, Meteer M, Van Ess-Dykema C 1998 Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*: 41, 443-492

Wright J, Carey M, Parris E 1995 Improved topic spotting through statistical modelling of keyword dependencies. In *Proceedings of the International Conference On Acoustics, Speech, And Signal Processing*, Detroit, pp 313-317.