Toward Understanding WH-Questions: A Statistical Analysis

Ingrid Zukerman[1] and Eric Horvitz[2]

[1] School of Computer Science and Software Engineering, Monash University, Clayton, Victoria 3800, AUSTRALIA, phone: +61 3 9905-5202, fax: +61 3 9905-5146, ingrid@csse.monash.edu.au

[2] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA, phone: +1 (425) 705-0917, fax: +1 (425) 936-7329, horvitz@microsoft.com

Abstract. We describe research centering on the statistical analysis of WH-questions. This work is motivated by the long-term goal of enhancing the performance of information retrieval systems. We identified informational goals associated with users' queries posed to an Internet resource, and built a statistical model which infers these informational goals from shallow linguistic features of user queries. This model was build by applying supervised machine learning techniques. The linguistic features were extracted from the queries and from the output of a natural language parser, and the high-level informational goals were identified by professional taggers.

Keywords: Statistical NLP, information retrieval, decision trees.

1 Introduction

The unprecedented information explosion associated with the evolution of the Internet makes salient the challenge of providing users with means for finding answers to queries targeted at large unstructured corpora. In addition to providing a large sea of heterogeneous information, the Web also provides opportunities for collecting and leveraging large amounts of user data. In this paper, we describe research on applying collaborative user modeling techniques to build models of users' informational goals from data gathered from logs of users' queries. The long-term aim of this project is to use these models to improve the performance of question-answering and information-retrieval systems. However, in this paper we focus on the user modeling component of this work.

We present modeling methods and results of a statistical analysis of questions posed to the Web-based Encarta encyclopedia service, focusing on complete questions phrased in English. We employ supervised learning to build a statistical model which infers a user's informational goals from linguistic features of the user's questions that can be obtained with a natural language parser. These informational goals are decomposed into (1) the type of information requested by the user (e.g., definition, value of an attribute, explanation for an event), (2) the topic, focal point and additional restrictions posed by the question, and (3) the level of detail of the answer. It is envisioned that our model of these informational goals will be used by different components of question-answering and information-retrieval systems. For instance, a document retrieval component could take advantage of the type of the requested information and the topic and focal point of a question; and an enhanced response generation system could additionally take into account the level of detail of the answer.

In the next section, we discuss related research. In Section 3, we describe the variables being modeled and our data collection efforts. In Section 4, we discuss our statistical model, followed by the evaluation of our model's performance. Finally, we summarize the contribution of this work and discuss directions for future research.

2 Related Research

Our research builds on insights obtained from using probabilistic models to understand free-text queries in search applications [Heckerman and Horvitz, 1998, Horvitz et al., 1998], and from the application of machine learning techniques to build predictive statistical user models.[1]

Previous work on statistical user models in IR includes the use of hand-crafted models and supervised learning to construct probabilistic user models that predict a user's informational goals. Heckerman and Horvitz (1998) and Horvitz *et al.* (1998) created Bayesian user models for inferring users' goals and needs for assistance in the context of consumer software applications. Heckerman and Horvitz' models considered words, phrases and linguistic structures (e.g., capitalization and definite and indefinite articles) appearing in free-text queries to a help system. Horvitz *et al.*'s models computed a probability distribution over a user's needs by considering the above linguistic parameters, a user's recent activity observed in his/her use of software, and probabilistic information maintained in a dynamically updated, persistent profile representing a user's competencies in a software application. Heckerman and Horvitz' models were used in a feature called *Answer Wizard* in the Microsoft Office'95 software suite. Horvitz *et al.*'s models were first deployed in the IR facility called *Office Assistant* in the Microsoft Office'97 office suite, and continue in service in the Microsoft Office 2000 package.

Lau and Horvitz (1999) built models for inferring a user's informational goals from his/her query-refinement behavior. In this work, Bayesian models were constructed from logs recorded by search services. These models relate the informational goals of users to the timing and nature of changes in adjacent queries posed to a search engine.

From an applications point of view, our research is most related to the IR arena of question answering (QA) technologies. QA research centers on the challenge of enhancing the response of search engines to a user's questions by returning precise answers rather than returning documents, which is the more common IR goal. Our work differs from QA research in its consideration of several user informational goals, some of which are aimed at supporting the generation of answers of varying level of detail as necessary. Further, in this paper we focus on the prediction of these goals, rather than on the provision of answers to users' questions. We hope that in the short term, the insights obtained from our work will assist QA researchers to fine tune the answers generated by their systems.

QA systems typically combine traditional IR statistical techniques with methods that might be referred to as "shallow" NLP. Usually, the IR methods are applied to retrieve documents relevant to a user's question, and the shallow NLP is

[1] For a survey of predictive statistical user models see [Zukerman and Albrecht, 2001].

used to extract features from both the user's question and the most promising retrieved documents. These features are then used to identify an answer within each document which best matches the user's question. This approach was adopted in [Kupiec, 1993,Abney et al., 2000,Cardie et al., 2000,Moldovan et al., 2000]. Abney *et al.* (2000) and Cardie *et al.* (2000) used statistical techniques centering on document and word frequency analysis [Salton and McGill, 1983] to perform document retrieval; while Kupiec (1993) and Moldovan *et al.* (2000) generated Boolean queries. Radev *et al.* (2000) and Srihari and Li (2000) adopted a different IR approach whereby the entities mentioned in documents are extracted first.

The NLP components of the above systems employed hand-crafted rules to infer the type of answer expected. These rules were built by considering the first word of a question as well as larger patterns of words identified in the question. For example, the question, *"How far is Mars?"* might be characterized as requiring a reply of type DISTANCE. In our work, we use supervised machine learning to build models that predict a user's informational goals from linguistic features of his/her questions. We seek to predict the type of the expected answer, its level of detail, and key aspects of its content.

3 Data Collection

Our models were built from questions identified in a log of Encarta Web queries. These questions include traditional *WH-questions*, which begin with "what", "when", "where", "which", "who", "why" and "how", as well as imperative statements starting with "name", "tell", "find", "define" and "describe". We extracted 97,640 questions (removing consecutive duplicates) out of a total of 1,649,404 queries logged by the WWW Encarta encyclopedia service during a period of three weeks in the year 2000. Thus, complete questions constituted approximately 6% of the total queries posed to this Web service. A total of 6,436 questions were tagged by hand. These questions had an average length of 6.63 words (compared to an average query length of 2.3 words in keyword-based queries [Lau and Horvitz, 1999]). Two types of tags were collected for each question: (1) tags describing linguistic features, and (2) tags describing attributes associated with high-level informational goals of users. The former were obtained automatically, while the latter were tagged manually.

We considered three classes of linguistic features: word-based features, structural features, and hybrid linguistic features.

Word-based features indicate the presence of specific words or phrases in a user's question, which we believed showed promise for predicting components of his/her informational goals. These are words like "make", "map", "picture" and "work".

Structural features include information obtained from an XML-encoded parse tree generated for each question by NLPWin [Heidorn, 1999] – a natural language parsing system developed by the Natural Language Processing Group at Microsoft Research. NLPWin analyzes queries, outputting a parse tree which contains information about the nature of and relationships among linguistic components, including parts of speech and logical forms. Parts of speech (PoS) include *adjectival phrases* (AJP), *adverbial phrases* (AVP), *noun phrases* (NP), *verb phrases* (VP), and *prepositional phrases* (PP). We extracted a total of 21 structural features including: the number of distinct PoS – NOUNs, VERBS, NPs, etc – in a question, whether the main noun is plural or singular, which noun (if any) is a proper noun, and the PoS of the head verb post-modifier.

Hybrid features are linguistic features constructed from structural and word-based information. Two hybrid features were extracted: (1) the type of head verb in a question, e.g., "know", "be" or `action verb`; and (2) the initial component of a question, which usually encompasses the first word or two of the question, e.g., "what", "when" or "how many", but for "how" may be followed by a PoS, e.g., "how ADVERB" or "how ADJECTIVE."

We considered the following variables representing high-level informational goals: *Information Need, Coverage Asked, Coverage Would Give, Topic, Focus, Restriction* and *LIST*. Information about the state of these variables was provided manually by three people, with the majority of the tagging being performed under contract by a professional outside the research team. To facilitate the tagging effort, we constructed a query-annotation tool.

Information Need is a variable representing the type of information requested by a user. We provided fourteen types of information need, including `Attribute, IDentification, Process, Intersection` and `Topic Itself` (which, as shown in Section 5, are the most common information needs), plus the additional category `OTHER`. As examples, the question *"What is a hurricane?"* was tagged as an `IDentification` query; *"What is the color of sand in the Kalahari?"* is an `Attribute` query (the attribute is "color"); *"How does lightning form?"* is a `Process` query; *"What are the biggest lakes in New Hampshire?"* is an `Intersection` query (a type of `IDentification` where the returned item must satisfy a particular *Restriction* – in this case "biggest"); and *"Where can I find a picture of a bay?"* is a `Topic Itself` query (interpreted as a request for accessing an object directly, rather than obtaining information about the object).

Coverage Asked and Coverage Would Give are variables representing the level of detail in answers. *Coverage Asked* is the level of detail of a direct answer to a user's question. *Coverage Would Give* is the level of detail that an information provider would include in a helpful answer. For instance, although the direct answer to the question *"When did Lincoln die?"* is a single date, a helpful information provider might add other details about Lincoln, e.g., that he was the sixteenth president of the United States, and that he was assassinated. The distinction between the requested level of detail and the provided level of detail makes it possible to model questions for which the preferred level of detail in a response differs from the detail requested by the user. We considered three levels of detail for both coverage variables: `Precise, Additional` and `Extended`, plus the additional category `OTHER`. `Precise` indicates that an exact answer has been requested, e.g., a name or date (this is the value of *Coverage Asked* in the above example); `Additional` refers to a level of detail characterized by a one-paragraph answer (this is the value of *Coverage Would Give* in the above example); and `Extended` indicates a longer, more detailed answer.

Topic, Focus and Restriction are variables that contain the PoS in the parse tree which represents the topic of discussion, the type of the expected answer and information that restricts this answer, respectively.[2] These variables take 46 possible values, e.g., $\mathtt{NOUN}_1$, $\mathtt{VERB}_3$ and $\mathtt{NP}_2$, plus the additional category OTHER. For each question, the tagger selected the most specific PoS that contains the portion of the question which best matches each of these informational goals. For instance, given the question "*What are the main traditional foods that Brazilians eat?*", the *Topic* is $\mathtt{NOUN}_2$ (*Brazilians*), the *Focus* is $\mathtt{ADJ}_3$+$\mathtt{NOUN}_1$ (*traditional foods*) and the restriction is $\mathtt{ADJ}_2$ (*main*). As shown in this example, it was sometimes necessary to assign more than one PoS to these target variables. At present, these composite assignments are classified as the category OTHER.

LIST is a Boolean variable which indicates whether the user is looking for a single answer (False) or multiple answers (True).

In addition, the tagger marked incoherent questions (BAD QUERY) and parse trees which did not match the user's question (WRONG PARSE). Also, the tagger entered clues from the questions which were helpful in determining *Information Need*, both types of *Coverage* and *LIST*. These clues formed the basis for linguistic features which were subsequently extracted automatically from questions. For instance, plural quantifiers such as "some" and "all" often indicate that a *LIST* of items is being requested.

4 Predictive Model

We built decision trees to infer high-level informational goals from the linguistic features of users' questions. One decision tree was constructed for each goal: *Information Need, Coverage Asked, Coverage Would Give, Topic, Focus, Restriction* and *LIST*. Our models were built using dprog [Wallace and Patrick, 1993], a procedure for constructing decision trees which is based on the Minimum Message Length principle [Wallace and Boulton, 1968].

The decision trees described in this section are those obtained using a training set of 4617 "good" questions and a test set of 1291 questions (both good and bad). Good questions were those that were considered coherent by the tagger and for which the parser had produced an appropriate parse tree (i.e., questions which were not BAD QUERIES and did not have a WRONG PARSE).[3] Our trees are too large to be included in this paper. However, we describe here the main attributes identified in each decision tree. For each target variable, Table 1 shows the size of the decision tree (in number of nodes) and its maximum depth, the attribute used for the first split, and the attributes used for the second splits. Table 2 shows examples and descriptions of the attributes in Table 1.[4]

We note that the *Focus* decision tree splits first on the initial component of a question, e.g., "how ADJECTIVE", "where" or "what", and that one of the second-split attributes

[2] Our *Focus* resembles the answer-type category considered by Kupiec (1993), Abney *et al.* (2000), Cardie *et al.* (2000) and Moldovan *et al.* (2000).

[3] The performance obtained with a larger training set comprised of 5145 queries (both good and bad) is similar to the performance obtained with this set.

[4] The meaning of "Total PRONOUNS" is peculiar in our context, because NLPWin tags words such as "what" and "who" as PRONOUNs. Also, the clue attributes, e.g., Comparison clues, represent convenient groupings of different clues that at design time were considered helpful in identifying certain target variables. These groupings reduce the number of attributes considered when building decision trees.

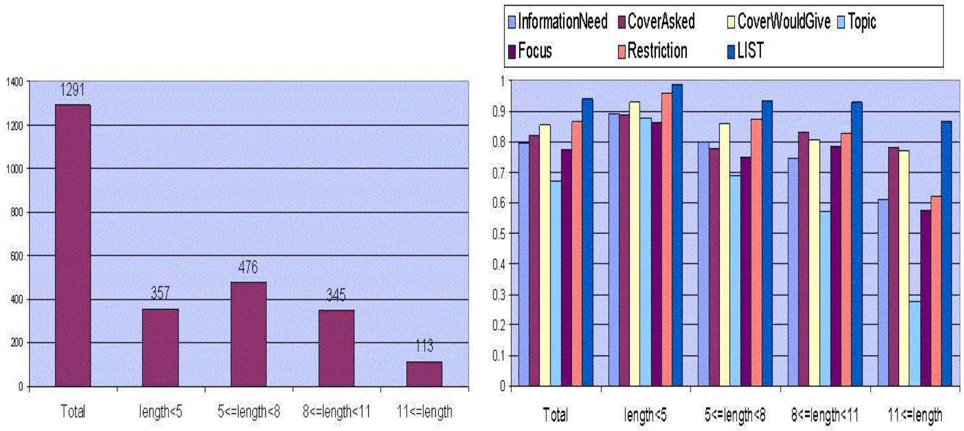
Table 1. Summary of decision trees

Target Variable	Nodes/Depth	First Split	Second Split
Information Need	207/13	`Initial component`	`Attribute clues, Comparison clues, Topic Itself clues, PoS after Initial component, verb-post-modifier PoS, Length in words`
Coverage Asked	123/11	`Initial component`	`Topic Itself clues, PoS after Initial component, Head verb`
Coverage Would Give	69/6	`Topic Itself clues`	`Initial component, Attribute clues`
Topic	193/9	`Total NOUNs`	`Total ADJs, Total AJPs, Total PRONOUNs`
Focus	226/10	`Initial component`	`Topic Itself clues, Total NOUNs, Total VERBs, Total PRONOUNs, Total VPs, Head verb, PoS after Initial component`
Restriction	126/9	`Total PPs`	`Intersection clues, PoS after Initial component, Definite article in First NP?, Length in phrases`
LIST	45/7	`First NP plural?`	`Plural quantifier?, Initial component`

Table 2. Attributes in the decision trees

Attribute	Example/Meaning
`Attribute clues`	e.g., "name", "type of", "called"
`Comparison clues`	e.g., "similar", "differ", "relate"
`Intersection clues`	superlative ADJ, ordinal ADJ, relative clause
`Topic Itself clues`	e.g., "show", "picture", "map"
`PoS after Initial component`	e.g., `NOUN` for "which country is the largest?"
`verb-post-modifier PoS`	e.g., `NP without PP` for "what is a choreographer"
`Total` *PoS*	number of occurrences of *PoS* in a question, e.g., `Total NOUNs`
`First NP plural?`	Boolean attribute
`Definite article in First NP?`	Boolean attribute
`Plural quantifier?`	Boolean attribute
`Length in words`	number of words in a question
`Length in phrases`	number of NPs + PPs + VPs in a question

is the PoS following the initial component. These attributes were also used to build the hand-crafted rules employed by the QA systems described in Section 2, which concentrate on determining the type of the expected answer (which is similar to our *Focus*). However, our *Focus* decision tree considers several additional attributes in its second split (these attributes are added by `dprog` because they improve predictive performance on the training data).



(a) Question length distribution (b) Predictive performance by question length

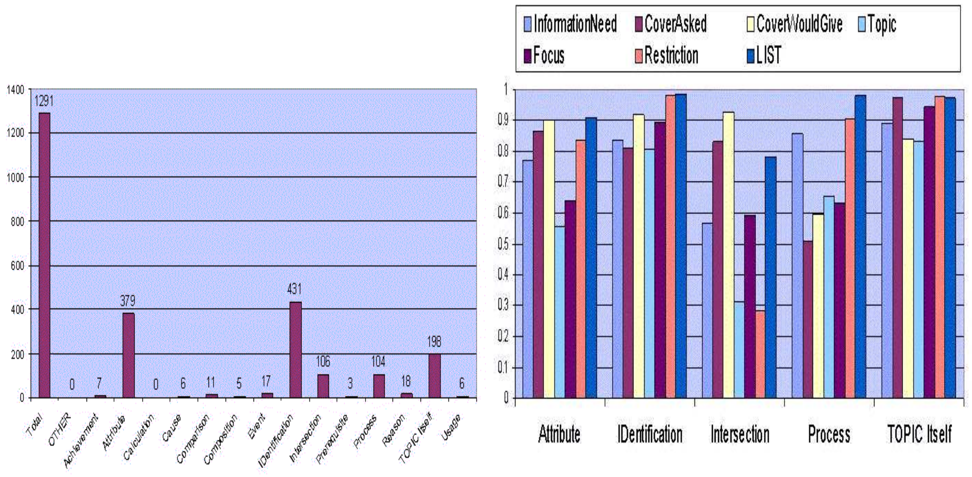
Fig. 1. Effect of question length on predictive performance

5 Results

We examine the effect of two factors on the predictive performance of our models: (1) question length (measured in number of words), and (2) information need (as recorded by the tagger).

Question length. The questions were divided into four length categories: less than 5 words, between 5 and 7 words, between 8 and 10 words, and more than 10 words. Figure 1(a) displays the distribution of questions in the test set according to these length categories. According to this distribution, over 90% of the questions have less than 11 words. The predictive performance of our decision trees broken down by question length is shown in Figure 1(b). As shown in this chart, for all target variables there is a downward trend in predictive accuracy as question length increases. Still, for questions of less than 11 words and all target variables except *Topic*, the predictive accuracy remains over 74%. In contrast, the *Topic* predictions drop from 88% (for questions of less than 5 words) to 57% (for questions of 8, 9 or 10 words). Further, the predictive accuracy for *Information Need, Topic, Focus* and *Restriction* drops substantially for questions that have 11 words or more. This reduction posts a "usability boundary" for the techniques proposed in this paper.

Information need. Figure 2(a) displays the distribution of the queries in the test set according to *Information Need*. The five most common *Information Need* categories are: `IDentification`, `Attribute`, `Topic Itself`, `Intersection` and `Process`, jointly accounting for over 94% of the queries. Figure 2(b) displays the predictive performance of our models for these five categories. The best performance is exhibited for the `IDentification` and `Topic Itself` queries. In contrast, the lowest predictive accuracy was obtained for the *Information Need, Topic* and *Restriction* of `Intersection` queries. This can be explained by the observation that `Intersection` queries



(a) Information need distribution

(b) Predictive performance for five most frequent information needs

Fig. 2. Effect of information need on predictive performance

tend to be the longest queries (as seen above, predictive accuracy drops for long queries). The relatively low predictive accuracy obtained for both types of *Coverage* for `Process` queries remains to be explained.

6 Discussion and Future Work

We have introduced a predictive model which can be used to infer key informational goals of a user from free-text questions posed to an Internet resource. The particular goals we have considered are: the user's information need, the level of detail requested by the user, the level of detail deemed appropriate by an information provider, and the topic, focus and restrictions of the user's question. The predictive model was constructed using a supervised machine learning technique under the collaborative approach. The performance of our model is encouraging, in particular for shorter queries and queries with certain information needs. However, further improvements are required in order to make this model practically applicable.

We believe there is opportunity to identify additional linguistic distinctions that could enhance the model's predictive performance. For example, we intend to represent frequent combinations of PoS, such as $\texttt{NOUN}_1\texttt{+NOUN}_2$, which are currently classified as `OTHER` (Section 3). We also propose to investigate predictive models which return more informative predictions than those returned by our current model, e.g., a distribution of the probable informational goals, instead of a single goal. This would enable an enhanced QA system to apply a decision procedure in order to determine a course of action. For example, if the `Additional` value of the *Coverage Would Give* variable has a relatively high probability, the system could consider more than one *Information Need, Topic* or *Focus* when generating its reply.

Our use of decision trees implicitly assumes independence between the variables that represent the different informational goals. However, this is not the case in reality. For instance, once a particular PoS is selected as the *Topic* of a question, it can no longer be its *Focus*. Likewise, *Information Need* influences both types of *Coverage*. In recent experiments we circumvented this problem to a certain extent by building decision trees which incorporate predicted values of informational goals. Our results indicate that it is worth exploring the relationships between several informational goals, with *Information Need* being a pivotal variable. We intend to use the insights obtained from this experiment to construct Bayesian networks, which will also capture probabilistic dependencies among these variables.

Finally, as indicated in Section 1, this project is part of a larger effort centered on improving a user's experience when accessing information from large information spaces. The next stage of this project involves using the predictions generated by our model to enhance the performance of QA or IR systems. One such enhancement pertains to query reformulation, whereby the inferred informational goals can be used to reformulate or expand queries in a manner that increases the likelihood of returning appropriate answers. As an example of query expansion, if *Process* was identified as the user's *Information Need*, words that boost responses to searches for information relating to processes could be added to the user's query prior to submitting it to a search engine. Another envisioned enhancement would attempt to improve the initial recall of the document retrieval process by submitting queries which contain the content words in the *Topic* and *Focus* of a user's question (instead of including all the content words in the question). In the longer term, we plan to explore the use of *Coverage* results to enable an enhanced QA system to compose an appropriate answer from information found in the retrieved documents.

Acknowledgments

This research was largely performed during the first author's visit at Microsoft Research. The authors thank Heidi Lindborg, Mo Corston-Oliver and Debbie Zukerman for their contribution to the tagging effort.

References

[Abney et al., 2000] Abney, S., Collins, M., and Singhal, A. (2000). Answer extraction. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pages 296–301, Seattle, Washington.

[Cardie et al., 2000] Cardie, C., Ng, V., Pierce, D., and Buckley, C. (2000). Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pages 180–187, Seattle, Washington.

[Heckerman and Horvitz, 1998] Heckerman, D. and Horvitz, E. (1998). Inferring informational goals from free-text queries: A Bayesian approach. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 230–237, Madison, Wisconsin.

[Heidorn, 1999] Heidorn, G. (1999). Intelligent writing assistance. In *A Handbook of Natural Language Processing Techniques*. Marcel Dekker.

[Horvitz et al., 1998] Horvitz, E., Breese, J., Heckerman, D., Hovel, D., and Rommelse, K. (1998). The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 256–265, Madison, Wisconsin.

[Kupiec, 1993] Kupiec, J. (1993). MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 181–190, Pittsburgh, Pennsylvania.

[Lau and Horvitz, 1999] Lau, T. and Horvitz, E. (1999). Patterns of search: Analyzing and modeling Web query refinement. In *UM99 – Proceedings of the Seventh International Conference on User Modeling*, pages 119–128, Banff, Canada.

[Moldovan et al., 2000] Moldovan, D., Girju, R., and Rus, V. (2000). Domain-specific knowledge acquisition from text. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pages 268–275, Seattle, Washington.

[Radev et al., 2000] Radev, D., Prager, J., and Samn, V. (2000). Ranking suspected answers to natural language questions using predictive annotation. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pages 150–157, Seattle, Washington.

[Salton and McGill, 1983] Salton, G. and McGill, M. (1983). *An Introduction to Modern Information Retrieval*. McGraw Hill.

[Srihari and Li, 2000] Srihari, R. and Li, W. (2000). A question answering system supported by information extraction. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pages 166–172, Seattle, Washington.

[Wallace and Boulton, 1968] Wallace, C. and Boulton, D. (1968). An information measure for classification. *The Computer Journal*, 11:185–194.

[Wallace and Patrick, 1993] Wallace, C. and Patrick, J. (1993). Coding decision trees. *Machine Learning*, 11:7–22.

[Zukerman and Albrecht, 2001] Zukerman, I. and Albrecht, D. W. (2001). Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):5–18.