When is a query a question? Reconstructing wh-requests from ad hoc-queries

Maarten van der Heijden
Dept. Computer Science,
Radboud University Nijmegen
m.vanderheijden@cs.ru.nl

Max Hinne
Dept. Computer Science,
Radboud University Nijmegen
mhinne@sci.ru.nl

Suzan Verberne
Centre for Language and
Speech Technology,
Radboud University Nijmegen
s.verberne@let.ru.nl

Eduard Hoenkamp
Dept. Knowledge Engineering,
Maastricht University
hoenkamp@acm.org

Theo van der Weide
Dept. Computer Science,
Radboud University Nijmegen
tvdw@cs.ru.nl

Wessel Kraaij
Dept. Computer Science,
Radboud University Nijmegen
TNO, Delft
kraaijw@acm.org

ABSTRACT

Search engine queries are often very short and usually lack explicit semantic structure or indication of intent. Recognizing intent is receiving increased attention since it enables a search engine to trigger a dedicated answer and presentation module. This way search engines increasingly move from a document ranking system to an answer facility. In this study we investigate whether it is possible to infer a hidden *wh*-request from a subset of all two- and three-word queries, based on the syntactic form of the queries. By analyzing dependency relations between the terms in the queries we gain insight in the structure of queries that are likely to have a *wh*-question as underlying intent. The goal of reconstructing a question from a keyword query is to be able to provide the user with an answer to that question, instead of only returning a list of documents.

1. INTRODUCTION

The first time people use a search engine they tend to type in a query the way they would ask information from a fellow human being. Very soon they learn that it is more efficient, quicker, and perhaps even more effective to just type in two or three keywords. There are a number of cases where the gain in speed is offset by the imprecision of the results. This paper focuses on ways to avoid the annoyance of imprecision by reconstructing, on the user's behalf, the question underlying the query.

Since web search engines such as Google and Bing have originally been designed as (web) document retrieval engines, queries are interpreted as ad hoc collections of words for which the most relevant web pages are found and ranked. The set of retrieved web pages is presented to the user as a ranked list of pointers to these documents consisting of their titles, URLs and descriptive snippets.

In recent years, Google and Bing have started to change their presentation of search results in accordance with the guessed intent of the users' queries. For example, for the query "restaurants amsterdam", Google first returns a map of the localized results, followed by the classic list of web pages. Similarly, for the query "lemur", Google presents a set of images and for "the ghost writer trailer" it returns a small set of YouTube-movies on top of the result list.

For some frequent *wh*-queries (questions starting with a *wh*-word such as *who* or *what*) Google provides database-like results on top of the result list. For example, the query "what is the capital of the netherlands" yields as first result the web definition for the word Amsterdam.[1]

Wh-questions are more specific than ad hoc-queries: Not only is the user's information need expressed more precisely by a natural language question than by the set of keywords typically used in an ad hoc-query, the unit of retrieval is also smaller and can be pointed out more specifically than the retrieval unit for ad hoc-queries. For example, the question "what is the capital of the netherlands" describes the searcher's information need very clearly and expects a clearly defined answer: the name of a city. If a user asks for a specific type of information by querying a *wh*-question, the search engine can help the user by providing an exact answer (in context) on top of the result list. However, only a fraction of queries has the form of a *wh*-question. Although Google provides the right answer for the query "capital netherlands", a more generic way to reconstruct query intent would allow answering less obvious questions.

In this paper, we assume that the proportion of queries that have the intent to be a *wh*-question is larger than the small proportion of fully formulated *wh*-questions, and we hypothesise that it is possible to predict the most likely *wh*-type from a two- or three-word query with a *wh*-intent. By predicting the type of *wh*-question that is intended by the user, a search engine can help the user by presenting a *wh*-answer at the top of the result list.

[1] In 2008 and 2009, Google used to provide database-like answers for more types of *wh*-questions. This functionality has largely been removed from the interface again.

2. RELATED WORK

Recognising the user's intent of search engine queries is a topic that has received only moderate attention in literature. In the traditional search engine approach, it is assumed that users' queries are informational/explorational in nature. The first study that shows that search engine queries are not always concerned with informational search is by Broder in 2002 [2]. Broder distinguishes three intent categories: navigational, informational and transactional. The main difference between the three categories seems to be more on a pragmatic level (i.e. related to the user's task) than related to the semantic interpretation level. Automatic classification techniques have been developed that can classify (most) queries in one of the three categories, in order to provide optimal search results [7].

Several authors use the term 'intent' as a more general notion of semantic interpretation. In the case of underspecified queries, underlying query intents can be guessed by suggesting possible query reformulations based on click information [4, 11]. Li et al. [9] study the automatic recognition of specific information needs (jobs, products), which they call 'job intent' and 'product intent'. Their methodology is based on exploiting the query-document click graph. Suggesting query reformulations has also been studied by looking at queries in semantically similar search *sessions* of other users [3]. Azzopardi and de Rijke [1] use an even broader notion of intent. They define intent as a condition on relevance: the underlying intent of the query determines which retrieval strategy is the most likely to give relevant results.

All interpretations and techniques in the literature share the basic intuition that many queries are underspecified and therefore the response of a search engine will often not be optimal. Recognition of possible user intents i.e. interpretations of the query on syntactic, semantic and pragmatic level is therefore a highly interesting research direction.

One query category that has a very explicit user intent, is the class of *wh*-questions: queries starting with *who*, *what*, *where*, *which*, *when*, *why* or *how*. By starting a query with a *wh*-word, the user explicitly mentions the type of answer he is searching for: a person in the case of *who*, a date or year in the case of *when*. In the literature on automated question answering (QA), the expected answer type is directly deduced from the *wh*-word in the question [6, 5, 10].

When we look at the queries in the log files of search engines, we see an important problem for the traditional QA approach that relies on the *wh*-word to predict the type of factoid to be retrieved: The proportion of queries that begin with a *wh*-word is very small: only 0.7% of queries in the Microsoft 2006 RFP dataset (a set of 14 million queries from US users entered into the Microsoft search engine in the spring of 2006) begin with a *wh*-word.[2] We assume that there is a large set of user queries that are in fact hidden *wh*-requests. These queries are underspecified and the correct recognition of the user's intent can help the search engine to provide the most relevant results.

3. RECOGNIZING THE TYPE OF IMPLICIT WH-REQUESTS

Contrary to database search, query languages for textual data are unstructured. In search engines that are intended for document retrieval, the query is represented as a bag of words, which is a very effective representation mechanism for document content. The bag-of-words model assumes that queries are unstructured and word order is irrelevant. However, we believe that queries do have some internal structure. From the specific combination of words in a query, the intended type of request may be derived. In some cases the intended type is more obvious than in others. For example, the query "make a blog", taken from Microsoft's click data (see Section 4), consists of a verb ('to make') with its direct object ('a blog'), and we can expect the user to have intended to get an answer to the request "how to make a blog".

In this paper we focus on hidden *wh*-requests. In an implicit *wh*-requests, the *wh*-word is not present while the user expects as result the answer to a specific *wh*-question. From the possible intents and their likelihoods, for each query a probability distribution over the *wh*-intent types (*who*, *how*, etc) can be derived. This probability distribution is denoted as $P(i \mid q)$ where i is a *wh*-intent type and q a query. The ambiguity of the query is defined as the entropy of this distribution function:

$$A(q) = -\sum_i P(i \mid q) \log P(i \mid q) \,. \tag{1}$$

In our experiments we have measured query type ambiguities to validate our assumption that there are useful query patterns for which the searcher's intent can be revealed with high certainty. In Figure 1 a histogram is shown that gives the number of queries binned over ambiguity (according to Equation 1). Each bin is simply defined as one tenth of the total entropy range. Although there are numerous queries for which the intent is difficult to reconstruct, at the lower end there is still a reasonable number of queries which would benefit from direct answers to the user's intended questions.

In Section 4, we use syntactic dependency parsing to reveal the most likely relation between query terms. We investigate the association between the type of dependency relation(s) that exist between the query terms and the likelihood of the query intent.

4. EXPERIMENT

In Section 1, we hypothesised that it is possible to predict the most likely *wh*-type from two- or three-word queries with a *wh*-intent. We used queries from the Microsoft 2006 RFP dataset to validate this hypothesis. This dataset consists of approximately 14 million queries from US users entered into the Microsoft Live search engine in the spring of 2006. For each query a number of details are available such as the document that was clicked on and its position in the list of results. We disregarded this information for the current experiment and only use the queries themselves.

Before we describe our method in technical detail, we first use an example to explain the general idea of our strategy.

4.1 Our strategy explained by example

The two-word query 'paper mache' has 47 occurrences in the click data. On top of that, the phrase 'paper mache'

[2] Another problem with the traditional QA approach is that many *wh*-questions are not actually factoid requests: the user does not expect a date or year as answer to the question "when should I divorce". However, this (interesting) problem lies outside the scope of this paper.

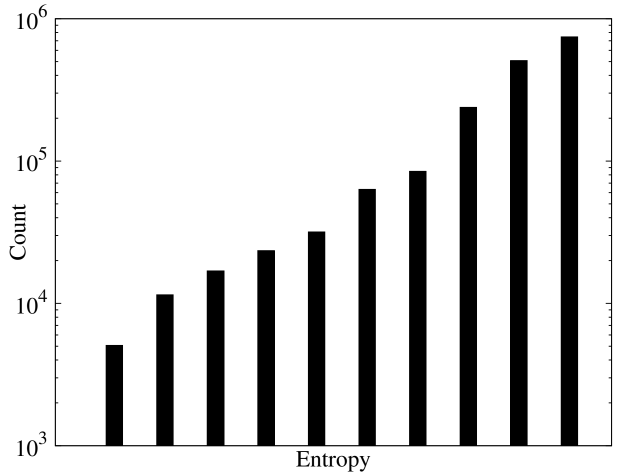

Figure 1: Histogram of the number of queries with a certain ambiguity (entropy). Each bin is 1/10 of the maximum entropy.

Table 1: Example: queries in which the query 'paper mache' is embedded with their frequencies

47	Number of occurrences of the query 'paper mache'
138	Queries in which the phrase 'paper mache' is embedded
54	Wh-queries in which the phrase 'paper mache' is embedded
25	how to make paper mache
8	how to paper mache
4	how to make paper mache masks
3	how to make a volcano paper mache
3	how to make a paper mache
2	how to paper mache rocks
2	how to paper mache for kids
2	how to make paper mache rocks
2	how to make a paper mache moon
2	how do I make a paper mache?
1	how to make paper mache glue

is embedded in 138 longer queries. 54 of these start with a *wh*-word. These numbers, and the *wh*-queries that contain the phrase 'paper mache', are shown in Table 1.

From these counts, we can conclude that of all 185 (138+47) queries that contain the phrase 'paper mache', 54 are *wh*-questions, and that all *wh*-queries that contain the phrase 'paper mache' are *how*-questions. This gives 'paper mache' a probability of $54/185 = 0.29$ that it has the intent 'how': $P(\mathit{how} \mid \mathit{paper\ mache}) = 0.29$. The other intents all have a probability of 0 for this query[3].

The dependency analysis of the phrase 'paper mache', according to the Connexor CFG parser [8], is:

```
1   paper   paper   attr:>2 @A> %>N N NOM SG
2   mache   mache   main:>0 @NH %NH N NOM SG
```

In the remainder of this paper, we represent a dependency relation between two words as `[word,rel,word]`, in this case `[paper,attr,mache]`, for the sake of readability.

By extracting the query counts and the dependency structure for all two- and three-word queries, we can infer general statistics about the association between within-query dependency relations (in this example `attr`) and the most likely *wh*-intent (in this example *how*).

Table 2: The most frequent dependency relations in the two- and three-word queries, according to the Connexor Parser.

Rel	Freq	Example
attribute	16 335	graduation speech [graduation,attr,speech]
modifier	1 362	set free [free,mod,set]
determiner	986	the 1800s [the,det,1800s]
object	855	make a movie [a,det,movie] [movie,obj,make]
preposition compl.	720	wall of china [of,mod,wall] [china,pcomp,of]
subject[4]	503	pineapples grow [pineapples,subj,grow]
rest	556	
total	21 317	

4.2 Method

Probability of query intent

We applied the strategy exemplified above to queries in the Microsoft 2006 RFP dataset. First, we extracted all two- and three-word queries from the total set of 14 million queries: 1 735 242 unique queries, and saved their frequencies. We also extracted all queries that start with a *wh*-word (42 151 unique queries).

We then derived the most likely *wh*-intent of a two- or three-word query as follows. Let again $P(i \mid q)$ be the probability that query q has intent i, i.e. the fraction of records in the data for which query q appears in conjunction with the *wh*-word corresponding to the intent i. The *wh*-intent of q is then simply the *wh*-word with the highest probability:

$$\mathit{wh}\text{-intent}(q) = \begin{cases} \operatorname{argmax}_i P(i \mid q) & \text{if } P(i \mid q) > d, \\ \text{undefined} & \text{otherwise.} \end{cases} \qquad (2)$$

The parameter d is an optional threshold for the intent probability that can be used to ensure that the query and query intent co-occur a minimum number of times before associating the intent with the query.

If we run this procedure with $d = 0$, we get 17 859 two- and three-word queries with a *wh*-intent. With $d = 0.2$, this is only 7 015. We decided to continue with the data for $d = 0$. Note that *wh*-queries are infrequent, most queries are navigational or transactional [2]. Since we do not know whether the user had a *wh*-intent, the distribution can only give an estimate that we cannot evaluate at this point.

Dependency relations

We used the Connexor CFG parser [8] to generate dependency relations for the two- and three word queries in our data set that have a *wh*-intent according to Eq. (2). The parser gives one dependency relation for a two-word query, and two dependency relations for a three-word query. As explained in Section 4.1, we convert the output of the parser to the easily readable format `[word,rel,word]`.

Of the 17 859 two- and three-word queries that have a *wh*-intent, 17 564 get an analysis from the Connexor parser. 3 753 of these (the three-word queries) consist two dependency relations. The total number of relations in the output is 21 317. The most frequent relations are listed in Table 2, together with an example of each of them and their frequency.

[3] Intents with probability 0 have been smoothed for the calculation of ambiguity (see Eq. (1))

[4] Among the subject relations, there are unfortunately many parser errors: other relations that are incorrectly labeled as 'subj'.

Table 3: Kullback-Leibler divergence for dependency relations ($d = 0$) that occur at least in 100 two- or three-word queries.

Dependency relation	KL-divergence	No of queries
object	0.198	855
quantifier	0.072	137
coordinating conjunction	0.066	190
determiner	0.052	986
preposition compl.	0.035	720
subject	0.026	503
modifier	0.005	1362
attribute	0.000	16335

Table 4: The scores (see text) for the dependency relations that occur at least in 100 queries and their most likely intent.

Dependency relation	Most likely intent	$Q_r(\hat{i})/P(\hat{i})$
object	how	1.408
subject	how	1.006
attribute	how	0.999
mod	how	0.971
preposition compl.	how	0.871
determiner	how	0.829
quantifier	how	0.814
coordinating conjunction	how	0.810

Associating dependency relations with intents

The next step in our approach is to determine whether a dependency relation significantly favors a particular *wh*-type. In the set of two- and three word queries that have a *wh*-intent, the most frequent intent is *how* (52%), which corresponds to the relative frequency of *how*-questions among all literal *wh*-questions in the click data set. The second-biggest *wh*-intent is *what* (29%); all others are relatively small. We use this distribution of *wh*-intent types as a background model for calculating the association strength between dependency relations and *wh*-intent types.

The next step in our approach was to determine whether a dependency relation has a significantly different intent distribution than the aggregated distribution. To do so, we calculated the Kullback-Leibler divergence between the *wh*-type distribution for each dependency relation Q_r against the background model (the aggregated distribution) P:

$$D_{KL}(P \parallel Q_r) = \sum_i P(i) \log \frac{P(i)}{Q_r(i)} , \qquad (3)$$

with $Q_r(i)$ the fraction of queries with dependency relation r and wh-type i and $P(i)$ the fraction of queries with wh-type i. The Kullback-Leibler divergences for the dependency relations that occur in at least 100 two- and three-word queries are shown in Table 3.

As our final step, we first identified the most likely *wh*-intent for each of the dependency relations in Table 3. We then calculated how much more likely this intent was for this relation compared to our background model, i.e. $\frac{Q_r(\hat{i})}{P(\hat{i})}$. Here $Q_r(\hat{i})$ is the probability of the most probable intent for this relation and $P(\hat{i})$ its probability in the background model. The scores are listed in Table 4.

5. CONCLUSION

In Table 3 we see that none of the dependency relations show a large KL-divergence in comparison with the aggregated query collection. This is confirmed in Table 4 where we see that each of the dependency relations we derived has 'how' as its most likely query intent. This is the most likely intent for the aggregated query collection as well, which explains the low divergence. However, we observe that the 'obj' relation favors a 'how'-intent more strongly than our background model. This provides an indication that queries phrased in the shape of an 'obj'-relation are likely to benefit from an answer that explains how to perform a certain action. Consider again a query from our running example, phrased as 'make paper mache' (an 'obj'-relation). According to our results, a user might benefit from a direct answer explaining '*how* to make paper mache'.

Although more aspects of query structure need to be covered before our suggestions can be applied to search engines, our approach shows promising results in reconstructing query intent. In future research we aim to find more relations between syntactical query structure and *wh*-intent. Ultimately, this would lead to a system that is able to differentiate between generic search and QA, in particular when users are likely to have an information need that can be satisfied by factoids.

6. REFERENCES

[1] L. Azzopardi and M. De Rijke. Query intention acquisition: A case study on automatically inferring structured queries. In *Proceedings the Dutch-Belgium Information Retrieval Workshop (DIR)*, 2006.

[2] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[3] S. Cucerzan and E. Brill. Extracting semantically related queries by exploiting user session information. Technical report, Microsoft, 2005.

[4] H. Daume and E. Brill. Web search intent induction via automatic query reformulation. In *Human Language Technology Conference/North American chapter of the Association for Computational Linguistics (HTL/NAACL)*, 2004.

[5] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, and A. Vilnat. Finding an answer based on the recognition of the question focus. *NIST Special Publication*, pages 362–370, 2002.

[6] E. Hovy, L. Gerber, U. Hermjakob, C.-J. Lin, and D. Ravichandran. Toward Semantics-Based Answer Pinpointing. In *Proceedings of the DARPA Human Language Technology Conference (HLT)*, San Diego, CA, USA, 2001.

[7] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, 44(3):1251–1266, 2008.

[8] T. Jarvinen and P. Tapanainen. Towards an implementable dependency grammar. In *Proceedings of COLING-ACL*, volume 98, pages 1–10, 1998.

[9] X. Li, Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2008.

[10] S. Lytinen and N. Tomuro. The use of question types to match questions in FAQFinder. In *AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 46–53, 2002.

[11] J. Mostert and V. Hollink. Effects of Goal-Oriented Search Suggestions. In *Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2008)*, 2008.