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

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

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

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 a query the way they would ask information from a fellow human being. 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.

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.

1In 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 information 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. 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 $P(i | q)$ 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:

$$A(q) = -\sum_i P(i | q) \log P(i | q).$$ (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’
In the remainder of this paper, we represent a dependency relation between two words as \([\text{word}, \text{rel}, \text{word}]\), in this case \([\text{paper}, \text{attr}, \text{mache}]\), 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 \text{attr}) and the most likely \text{wh}-intent (in this example \text{how}).

3Intents with probability 0 have been smoothed for the calculation of ambiguity (see Eq. (1)).
Associating dependency relations with intents

In the set of two- and three-word queries that have a
dependency relation significantly favors a particular
model. The scores are listed in Table 4.

\[ D_{KL}(P \parallel Q_r) = \sum_i P(i) \log \frac{P(i)}{Q_r(i)} \]  

with \( Q_r(i) \) the fraction of queries with dependency relation \( r \) and \( P(i) \) the fraction of queries with \( w \)-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 \( w \)-
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. \( \frac{Q_r(i)}{P(i)} \).

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

<table>
<thead>
<tr>
<th>Dependency relation</th>
<th>KL-divergence</th>
<th>No of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>object</td>
<td>0.198</td>
<td>855</td>
</tr>
<tr>
<td>quantifier</td>
<td>0.072</td>
<td>137</td>
</tr>
<tr>
<td>coordinating conjunction</td>
<td>0.066</td>
<td>190</td>
</tr>
<tr>
<td>determiner</td>
<td>0.052</td>
<td>986</td>
</tr>
<tr>
<td>preposition compl.</td>
<td>0.035</td>
<td>720</td>
</tr>
<tr>
<td>subject</td>
<td>0.026</td>
<td>503</td>
</tr>
<tr>
<td>modifier</td>
<td>0.005</td>
<td>1362</td>
</tr>
<tr>
<td>attribute</td>
<td>0.000</td>
<td>16335</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dependency relation</th>
<th>Most likely intent</th>
<th>( Q_r(i)/P(i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>object</td>
<td>how</td>
<td>1.408</td>
</tr>
<tr>
<td>subject</td>
<td>how</td>
<td>1.006</td>
</tr>
<tr>
<td>attribute</td>
<td>how</td>
<td>0.999</td>
</tr>
<tr>
<td>mod</td>
<td>how</td>
<td>0.971</td>
</tr>
<tr>
<td>preposition compl.</td>
<td>how</td>
<td>0.871</td>
</tr>
<tr>
<td>determiner</td>
<td>how</td>
<td>0.829</td>
</tr>
<tr>
<td>quantifier</td>
<td>how</td>
<td>0.814</td>
</tr>
<tr>
<td>coordinating conjunction</td>
<td>how</td>
<td>0.810</td>
</tr>
</tbody>
</table>

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 ben­
cfit 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). Accord­
ing 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 cov­
ered before our suggestions can be applied to search en­
gines, our approach shows promising results in reconstruc­
ting 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 differ­
entiate between generic search and QA, in particular when
users are likely to have an information need that can be satis­
fied by factoids.

6. REFERENCES

  case study on automatically inferring structured queries. In
  Proceedings the Dutch-Belgian Information Retrieval Workshop
  (DIR), 2006.
  queries by exploiting user session information. Technical report,
  Microsoft, 2005.
  automatic query reformulation. In Human Language
  Technology Conference/North American chapter of the
  Association for Computational Linguistics (HLT/NAACL),
  2004.
  L. Monceaux, I. Robba, and A. Vilnat. Finding an answer
  based on the recognition of the question focus. MIST Special
  D. Ravichandran. Toward Semantics-Based Answer
  Pinpointing. In Proceedings of the DARPA Human Language
  Technology Conference (HLT), San Diego, CA, USA, 2001.
  informational, navigational, and transactional intent of Web
  queries. Information Processing and Management,
  dependency grammar. In Proceedings of COLING-ACL,
[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.
    questions in FAQFinder. In AAAI Spring Symposium on
    Mining Answers from Texts and Knowledge Bases, pages
    on Artificial Intelligence (BNAIC 2008), 2008.