Ambiguous Queries: Test Collections Need More Sense

Mark Sanderson
Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@shef.ac.uk

ABSTRACT

Although there are many papers examining ambiguity in Information Retrieval, this paper shows that there is a whole class of ambiguous word that past research has barely explored. It is shown that the class is more ambiguous than other word types and is commonly used in queries. The lack of test collections containing ambiguous queries is highlighted and a method for creating collections from existing resources is described. Tests using the new collection show the impact of query ambiguity on an IR system: it is shown that conventional systems are incapable of dealing effectively with such queries and that current assumptions about how to improve search effectiveness do not hold when searching on this common query type.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Systems and Software --- performance evaluation.

General Terms

Measurement, Experimentation.

Keywords

Ambiguous queries, test collections, diversity.

1. INTRODUCTION

Word sense ambiguity is a topic that has been studied for many years in the Information Retrieval (IR) community, starting with Weiss's small scale experiments [21] through to a more thorough examination of the topic in the 1990s. Most of the past disambiguation research focussed on ambiguity of words found in dictionaries, which have poor coverage of proper nouns or phrases such as titles, names, etc. This is unfortunate since it is increasingly clear that names of people, locations, organizations, acronyms, etc, are common queries in search engines. Some of these nouns will have high levels of ambiguity, but the extent of the ambiguity is little understood.

While disambiguation research was studied explicitly, it was also studied implicitly with research on result list clustering, sub-topic retrieval and other algorithms for increasing *diversity* in search results. While some of this research has shown improvement in retrieval effectiveness, studies of this type are hampered by a lack of test collections containing ambiguous queries.

SIGIR'08, July 20–24, 2008, Singapore.

Consequently, this paper describes a re-examination of the problem of word ambiguity and an exploration of possible solutions to the lack of appropriate test collections for this area of IR. The work addresses the following research questions:

- How common is ambiguity in words not normally found in conventional dictionaries?
- What form does such ambiguity take?
- How common is ambiguity in queries to search engines?
- How well do existing test collections support research in ambiguous queries?
- How much does ambiguity impact on IR effectiveness?

This paper describes the study, starting with a review of past literature. The main data set used in the ambiguity study is described next followed by an analysis of it. Next, query logs are examined for ambiguity. This is followed by a study of ambiguity on a search engine using a variety of ranking algorithms.

2. LITERATURE REVIEW

In this Section, the past work in the following topics is examined: research on the extent and nature of ambiguity when searching; research seeking to diversify the documents appearing in a ranking; research on word sense disambiguation and the limited work on building test collections with ambiguous queries.

2.1 The extent of ambiguity

Krovetz [13] conducted a careful examination of the extent of word sense ambiguity in test collections (e.g. TIME and CACM). He found that the different senses of ambiguous query words provided good separation between relevant and non-relevant documents. In addition, by looking up query words in a dictionary (Longman's Dictionary of Contemporary English) he was able to calculate the average number of senses per non-stop word in the test collection topics. For CACM it was 5.3, for TIME it was 4.8. However, he found that retrieval based on the queries, rarely benefited from disambiguation as highly ranked documents tended to match on a number of words in the query. In such situations, ambiguous query words generally match only on the correct sense. For example, despite the word "bat" being ambiguous, the query "bat echolocation" is unlikely to retrieve top ranked documents referring a sporting implement.

At the time of Krovetz's study, it was largely assumed that queries to ranked retrieval systems would typically be sentence like statements expressing detailed information needs. Jansen and Spink amongst others showed that queries are typically much shorter [12], therefore potentially more ambiguous. In addition, Krovetz's work along with other disambiguation research of the time largely focussed on ambiguity as defined in dictionaries or online thesauri such as WordNet. Such reference corpora provided excellent coverage of ambiguous words. However, their coverage of proper nouns was poor.

2.1.1 Examination of proper nouns

In recent years there has been a growth of research into certain types of ambiguous proper nouns with particular focus on people and place names. (Note, the different entities an ambiguous proper noun could refer to are known as *referents*.)

SemEval 2007 ran the Web People Search evaluation (WePS, Artiles, et al, [2]), which focussed on the disambiguation of individuals when their names are searched for on a retrieval system. Artilles et al justified their investigation by stating that the United States Census Bureau record that 90,000 names are shared by 100 million people in the U.S. A training corpus composed of 100 web pages, retrieved by the Yahoo! search engine, for each of 39 person names was built; each page was manually annotated to mark up the different people (i.e. the referents) sharing a name. A test set of the 100 pages retrieved for each of 30 further person names was also created and annotated. Seventeen participants submitted runs that mostly outperformed a baseline. The authors of the evaluation campaign noted, however, that there were some unexpected qualities to the test data and it is possible that this may have had an impact on this year's results.

There has also been much work examining the task of disambiguation of place names. Leidner provides a thorough survey of this research field [14]. Unlike WePS, there is as yet no common evaluation test bed for comparing the effectiveness of place name disambiguation systems.

What is currently missing from work in these two research areas is a study of the extent of such ambiguity in user queries and any impact it might have on retrieval effectiveness.

2.2 Disambiguation and IR

Voorhees [20] conducted the first large scale study of a word sense disambiguation system applied to the topics and documents of 5 test collections. Words were disambiguated and the retrieval effectiveness of an IR system applied to those collections was compared to the effectiveness of the system searching on the collection without disambiguation. Perhaps unexpectedly, use of disambiguation reduced effectiveness. Analysing Voorhees' result, Sanderson [15] used a process for simulating ambiguity called *pseudo-words*[1]. He presented results showing the negative impact of disambiguation errors on retrieval effectiveness and suggested that this was the cause of Voorhees's negative result. He also showed how retrieval based on very short queries was affected by ambiguity.

Researchers continued to study the topic and in later work, Sanderson et al showed some limited value in using word sense disambiguation in retrieval [16], as did Gonzalo et al [11]. Both studies, however, were on small or very small test collections. In the field of cross language retrieval, correct query translation presupposes some form of ambiguity resolution of query words. Work from Darwish and Oard [8], showed success through consideration of sense. In monolingual research, Stokoe et al [19] showed a clear value in conducting disambiguation using one of the relatively large TREC Web corpora, WT10G.

All of the research described in this Section makes a common assumption: that at the time an ambiguous query is submitted to a search engine, it will be known somehow, which sense of each ambiguous word was intended by the person who issued the query. Such information might be worked out from the query itself (if the query is detailed enough), from a profile of the user held by the search engine, from information on the context of the search, click data from past searches, or simply by asking the user to clarify their information need.

However, it is not clear that such information will always be available to an operational search engine. If such data is missing, the only alternative is for the engine to return a diverse ranking composed of documents that are relevant to different interpretations of the ambiguous query. This appears to be a reasonable strategy in the face of such queries, however, there are no publicly available test collections that contain ambiguous queries in order to test such a strategy.

2.3 Test collections: 1 interpretation per topic

Across virtually all test collections used in IR research, topics have a single interpretation, which is explicitly defined in the topic's description and/or narrative and implicitly defined in its relevance judgements. Collections are set up in this way as each topic is created to represent the information need of one person.

The idea of defining single interpretations for test collection topics started with Cleverdon when building the Cranfield collection. In a reflective piece [7] he described how this feature of Cranfield was introduced after he witnessed the problems of two search groups in the 1950s who attempted to compare their retrieval systems. The groups agreed on a common set of documents and topics for testing. However, the relevance judgements, and therefore the interpretation of the topics, were left undefined. Cleverdon reported that when the two groups came together to compare effectiveness scores, they discovered they had different relevance judgements because they had interpreted the topics differently. Cleverdon recounts that the groups spent the whole of the first day of their meeting arguing about the interpretation of just the first topic.

One lesson Cleverdon appears to have drawn is the importance of defining relevance judgements in a test collection. The other lesson is to base judgements on one person's interpretation of the topic. While most would agree on the correctness of the first lesson, in the light of the short ill defined queries submitted to modern search engines, it is less clear how many would agree on the universality of the second.

2.4 Test collections with sub-topics

There is a collection that tested an area related to ambiguous queries: the 20 topic collection built for the TREC interactive track. Its topics addressed broad themes that had within them a number of *instances* or *sub-topics* that search engines were expected to retrieve. The instances were identified by TREC assessors: on average, there were 20 per topic. To illustrate, topic 353i (selected at random) from TREC 7 has the title "*Antarctic exploration*", the user need was to "*Identify systematic explorations and scientific investigations of Antarctica, current or planned*". In total, 11 instances were identified

- mining prospection
- oil resources
- rhodium exploration
- ozone hole / upper atmosphere
- greenhouse effect
- measuring chemicals in the atmosphere
- analysis of toxic waste
- whale scientific research
- antarctic sonar mapping
- ice studies
- climate studies

[1] Developed by both Gale et al [9] and by Schütze [17].

More recently, 20 topics in the Million Query Track of TREC (Allan et al, [1]) were also created with multiple instances: on average 3.1 per topic. Clarke et al [6] pointed out that the TREC 2005 and 2006 Question Answering collections have similarities to the instance/sub-topic collections and used the QA collections as a proxy for experiments with ambiguous queries.

2.5 Research promoting diverse ranks

It would appear that the approach used to ensure that different sub-topics of a general query are retrieved, is broadly the same as the approach used to ensure that different interpretations of an ambiguous query are retrieved: the approach being some form of clustering. Most of the early work on clustering (surveyed by Willett, [22]) focused on sub-topic retrieval. Maximal Marginal Relevance (MMR) from Carbonell et al [3] and more recently the work of Zhai et al [23] continued to study this area. Zhai et al used the 20 TREC interactive track topics to show the worth of their approach.

Chen & Dumais [4] reported that users locate relevant items quicker in search results that are organised into well defined clusters. The queries they tested on were ambiguous. Building search systems to tackle such queries was reported by Zhang et al [24] as well as Chen & Karger [5]. Zhang et al measured effectiveness using an in-house test collection; Chen and Karger found a number of creative ways to test their system using TREC data. However, they were essentially limited to testing sub-topic retrieval.

Without test collections containing ambiguous topics with associated relevance judgements that reflect a range of interpretations of that topic, the worth of much of the work described here may not be fully understood.

2.6 Lessons from past work

Although past work has examined many aspects of ambiguity, it is not clear that assumptions behind that work still hold; the assumptions being: queries are long; ambiguous words in queries are not proper nouns and a search engines will know the correct interpretation of a topic at retrieval time. What appears to be missing from the body of past research is an understanding of the extent to which proper nouns are ambiguous and the extent of ambiguity in user queries across all forms of ambiguous words. Finally, as described here and in past work [18], there is a lack of test collections for examining topic ambiguity.

3. A STUDY OF AMBIGUITY

To re-examine the impact of ambiguity on search engine effectiveness, it was necessary to locate sources of ambiguous words and phrases. The classic study on the extent of ambiguity in test collections from Krovetz acknowledged the dictionary he used did not cover proper nouns. Consequently, we sought sources that had better coverage of such word forms. WordNet (version 3.0), contains 87,633 words in the English language of which 16,882 (19%) are proper nouns[2]. An examination of a randomly selected sample of 1% of these nouns revealed a wide range of place names, scientific terms, names of plants, people, organizations, objects, etc.

There was a concern that as a source, WordNet did not cover a sufficient range of proper nouns: Artilles et al described 90,000 person names alone. Type specific lists of nouns and their referents, can be obtained, however, ambiguous words are likely to have referents across a number of types: Springfield for example is both the name of many towns and cities, while at the same time being the name of a fictitious place in a popular television program and a relatively common surname.

The online encyclopedia, Wikipedia, covers a great many topics that one might assume largely reflect the interests and information needs of the population of users that access many search engines. Of particular interest are the resource's great many so-called *disambiguation pages* that list the referents of ambiguous words and phrases most of which are proper nouns. The complete data set can be downloaded and the vast majority of pages relating to ambiguity can be identified relatively easily, by either matching the string "_disambiguation" in the title of the article or, more commonly, finding the "{{disambig}}" template tag. The vast majority of disambiguation pages conform to an easily parsed format. Therefore, it was decided to use this collection to study ambiguity of Proper nouns. This was complemented with WordNet's lists of other ambiguous word forms.

3.1 Obtaining data sets

The 12.7Gb collection "enwiki-20071018-pages-articles.xml" – a snapshot of the Wikipedia English language pages without edit history from late 2007 – was downloaded. It contained ≈2.2 million articles. It was parsed for disambiguation pages. In total 69,769 pages, identified just by the "{{disambig}}" tag, were extracted. A further 29,504 pages were identified by the title string "_(disambiguation)". Of the later set of pages, 10,164 were removed as they were found to be pages that simply re-direct to other disambiguation pages, leaving 19,340 articles remaining. In total the number of ambiguous words and phrases in Wikipedia identified through these two page forms was 89,109.

The two page types are a widely used convention in Wikipedia to indicate different forms of ambiguity. If, through the collaborative processes of forming articles, it was "decided" that a word or phrase had clearly one main referent, then that referent would be described in an article entitled with that word or phrase. The other referents were then packaged into a disambiguation page with the string "_disambiguation" added at the end of the page title. For example, the article "Chicago" is about the large city in the United States, and the other referents of that word are listed in the page "Chicago_(disambiguation)".

If, however, there is no consensus on a dominant sense for a word or phrase, the page for that word will be a disambiguation page. For example, the word "Babcock" has many referents listed (e.g. person names, places, etc), no single referent of which is considered to be dominant.

The structure of disambiguation pages was found to be consistent. (Determined by examining 182 disambiguation pages randomly selected; of these, 4 (2%) failed to adhere to the format.) Each referent was displayed in a bulleted list and referents of the same type were grouped together with the group given a title (see Figure 1). As can be seen, there is value in referring to the different uses of "Babcock" as referents rather than senses, as the Wikipedia list of the range of ways in which this term can be used clearly are not senses in the traditional definition of the word. For this particular term, the word is most likely used as part of a longer name. Nevertheless, it is likely that a search engine could

[2] These were identified simply by capitalization.

People

- Alpheus Babcock (1785–1842) American piano and musical instrument maker
- Barbara Babcock (b. 1937) American actress
- Courtney Babcock (b. 1972) Canadian runner
- Edward V. Babcock (1864–1948) former mayor of Pittsburgh, PA
- E. B. Babcock (Ernest Brown Babcock, 1877–1954) American plant geneticist
- George Herman Babcock (1832–1893) American inventor
- Harold D. Babcock (1882–1968) American astronomer
- Horace W. Babcock (1912–2003) American astronomer
- Ira L. Babcock (1808-1888) American pioneer and judge
- John Babcock (b. 1900) Last surviving Canadian WWI veteran
- Joseph Park Babcock (1893-1949) American Mahjong promoter
- Mike Babcock (b. 1963) Canadian hockey coach & former player
- Orville E. Babcock (1835–1884) American Civil War General
- Rob Babcock, former general manager of the Toronto Raptors NBA basketball team
- Roscoe Lloyd Babcock, (1897-1981) California artist
- Stephen Moulton Babcock (1843–1931) American agricultural chemist
- Tim M. Babcock (b. 1919) former governor of Montana

Geographic places

- Babcock (crater), a lunar crater
- Babcock Lakes, former ponds in Washington, D.C.
- Babcock, Michigan, an unincorporated town
- Babcock State Park, a West Virginia state park

Science

- Babcock Model, a mechanism to describe sunspot patterns
- Babcock test, which determines the fat content of milk

History and law

- Babcock Amendment, Minnesota constitutional amendment

Corporations and schools

- Babcock and Brown, an Australian investment bank
- Babcock and Wilcox, an American company producing energy production and pollution control equipment
- Babcock Electric Carriage Company, early 20th century US automobile maker
- Babcock International Group, a British engineering and services company
- Babcock University, Seventh-Day Adventist university in Nigeria
- Babcock Graduate School of Management, Wake Forest University's business school, located in Winston-Salem, NC.

Fictional characters

- Iris Babcock, Sergeant-Major from the Honorverse
- C. C. Babcock, on the sitcom *The Nanny*

Figure 1. Example of a Wikipedia disambiguation page

be presented with such a single word query and in the absence of any context information; the engine would do well to retrieve relevant documents referring to a number of these referents.

WordNet v3.0 was downloaded and the word definition information stored in the thesaurus's data files for each of the four main grammatical forms were examined. In total, 87,633 words were listed in WordNet, of which 15,302 (17.5%) were found to be ambiguous. Using the heuristic that a proper noun can be identified by the word starting with a capital letter, 16,882 such nouns were found, of which 1,297 (7.7%) were ambiguous.

3.1.1 Data sets used

Three lists of ambiguous words/phrases were created: those from WordNet (labeled *WN*); those disambiguation pages in Wikipedia for which one referent was considered dominant (Wi_D); and those pages for which no dominant referent existed (Wi_{ND}). The lists of particular interest were *WN* and Wi_{ND}. The former as it represented the words used in past studies of ambiguity; the later because it represented terms that are most likely to be important in further studies of ambiguity.

4. STUDYING WIKIPEDIA

An examination of Wikipedia's disambiguation pages was conducted, examining the number of referents, the length of the ambiguous words or phrases and the types of referents.

4.1 Number of referents

Given the consistency of formatting of disambiguation pages, it was possible to determine the number of referents in the pages simply by counting the number of bullet points on the page[3]. Table 1 shows the results of the study.

Referents	Pages	%	Referents	Pages	%
2	23002	*26*	12	1473	*2*
3	14128	*16*	13	1293	*1*
4	9647	*11*	14	1180	*1*
5	7063	*8*	15	1001	*1*
6	5308	*6*	16	835	*1*
7	4108	*5*	17	754	*1*
8	3241	*4*	18	671	*1*
9	2769	*3*	19	607	*1*
10	2194	*3*	20	561	*1*
11	1849	*2*	>20	6308	*7*

Table 1. Number of referents in Wikipedia

Senses	Words	%	Senses	Words	%
2	9538	*62*	7	227	*1*
3	2845	*19*	8	126	*1*
4	1176	*8*	9	103	*1*
5	603	*4*	10	70	*0*
6	395	*3*	>10	219	*1*

Table 2. Number of senses in WordNet

The total number of referents in Wikipedia was approximately 617,000. Compared to the total number of English articles in Wikipedia, approximately 2.2 million, one can calculate that 28% of all articles in Wikipedia were referred to from a disambiguation page. As can be seen, the number of referents per page follows a power law distribution with a long tail: the number of pages with 10 or more referents was substantial, constituting 22% of all pages. The average number of referents was 7.39.

The distribution of referents was compared to the distribution of senses found in WordNet (shown in Table 2). As above, a power law distribution was observed, but with WordNet, the peak was higher and the tail was not as long. The average number of word senses was 2.96.

[3] The correctness of this simple heuristic was checked by analysing the random sample of 182 disambiguation pages from Wikipedia. It was found that counting bullet points over estimated the number of referents by under 5%. Note, it was occasionally found that a referent was further broken down into "sub-referents", these were ignored in this analysis.

4.2 Size of ambiguous words/phrases

The size of the ambiguous words/phrases was also calculated as shown in Table 3. As can be seen, the majority of ambiguous terms in Wikipedia were one word long, however many longer terms exist also. A random sample of terms with >4 words were examined and found to be often titles of songs, films, television programs that have been used by more than one production.

Len	Num	%	Len	Num	%
1	42,500	*61*	1	14157	93%
2	19,319	*28*	2	1077	7%
3	5,317	*8*	3	53	0%
4	1,799	*3*	4	10	0%
>4	834	*1*	>4	5	0%

Table 3. Word length of ambiguous terms in Wikipedia (left) and in WordNet (right)

Table 3 also details the word length of ambiguous terms in WordNet. A comparison was made between the two resources. As can be seen, the relative number of one word entries in WordNet is larger compared to the number in Wikipedia. This is potentially important as in past disambiguation research, it was assumed that ambiguity was largely restricted to single words. Multi-word queries, were assumed to be relatively unaffected by ambiguity. However, it appeared from this analysis that such queries are potentially ambiguous as well. The likelihood of this happening was tested and the results are shown in Sec. 5.2.

4.3 Referent types

Given that disambiguation pages often list referents grouped by type, the titles of each groups were extracted and analyzed to determine the referent types present in Wikipedia. The naming of the groups does not follow a strong convention. Therefore, some normalization was conducted: the 40 most used group names were identified and manually arranged into common types.

Type	%	Type	%
People	*27*	Science	*3*
Places	*26*	Ships	*1*
Other/Misc	*25*	Sports	*1*
Entertainment/Music/Films/TV	*11*	Military	*1*
Companies/Organizations	*5*		

Table 4. Types of referents

The relative percentages of the identified types are shown in Table 4. As can be seen, names of people and locations are the predominant single type. Names of televisions programs, films, music, etc were also relatively common with other types, less so. This analysis examined the fraction of referent types within disambiguation pages, what the analysis did not examine was the fraction of pages who's referents were entirely one type or another, this was the objective of the next study.

4.4 Mix of sense types per page

To conduct this analysis a manual examination of a random sample of 182 disambiguation pages (0.25% of all such pages) was conducted. It was found that 44% of the pages examined had no consistent referent type. The fraction of pages with referents that were exclusively the names of a person was 9%, the fraction that were just locations was 11%. From this analysis, it would appear that the work described in the literature review (Sec. 2.1.1) that primarily studies particular referent types in isolation, risks ignoring the heterogeneous referents of many ambiguous proper nouns.

4.5 Discussion

Many differences between WordNet and Wikipedia were identified, justifying the re-examination of the relationship between ambiguity and information retrieval effectiveness.

- ambiguous proper nouns are ambiguous across a number of different forms, e.g. person names, places, titles, etc, with no single type accounting for a majority of occurrences;
- the number of ambiguous words which have referents across different types is large;
- there are more ambiguous multi-word terms than has previously been described in the literature;
- the number of referents per ambiguous proper noun is larger than that found in the resources previously used is disambiguation research.

What was required next was a study of the extent to which such ambiguity exists in the queries submitted to search engines.

5. Analyzing search engine logs

The next stage of the analysis was to examine the prevalence of ambiguity in actual searches. To achieve this, the logs of a number of search engines were examined to determine the fraction of the logs that contained ambiguous words/phrases in the Wikipedia and in the WordNet lists. It has been established that ambiguous words within a longer multi-word queries tend not to cause reductions in retrieval effectiveness, therefore, only those queries that entirely matched an ambiguous word/phrase were considered in this analysis.

It was considered important to examine different queries, therefore (with the permission of the owners) logs from the following two search engines were obtained: a large Web search system from Microsoft (labeled Web) and a large journalistic photo library from the UK's Press Association (PA, searching between 10 and 20 million images).

Log	Unique queries (all)	Most frequent (fr)	Year(s) gathered
Web	1,000,000	8,719	2006
PA	507,914	14,541	2006-7

Table 5. Attributes of query logs

The logs were stripped of all user, session and click related data, leaving just the query text. Query words were reduced to lower case. The frequency of occurrence of each query in a log was counted to produce a histogram file, each line of which contained a query/frequency pair for each unique query in the log. Both the whole log (labeled *all*) and a sub set of the log containing the most frequent queries (labeled *freq*) were examined. The method for forming *freq* was to select those queries that occurred at least *n* times more often than the least frequent query. The value of *n* (87) was selected at random from a range of possible values that would create a small subset. The statistics for the log histograms are shown in Table 5.

As can be seen, the size of the logs is similar, as is the year(s) in which the logs were gathered. Note, at the request of the log owners, the exact time period over which the logs were gathered

was withheld in order not to reveal information about the traffic volumes. The lack of this detail was not judged to have an impact on the results of or the conclusions drawn from the experiments conducted here.

5.1 Number of ambiguous queries

The focus of interest in this experiment was in determining the fraction of queries that were listed as being ambiguous. The fraction was measured in the three lists and was measured across the two subsets of each of the logs. The figures detailing the measurements are shown in Table 6. As can be seen, even though the logs are from search engines serving different collections and different user needs, there are certain consistent trends. Examining each column in turn:

- Column 1 records the overlap with Wi_{ND}: a noticeable number of queries are ambiguous, particularly in *freq*.
- Column 2 records the generally smaller number of queries that are listed as ambiguous in WordNet (*WN*).
- Column 3 details the overlap between the query logs and the union of the *WN* and Wi_{ND} combined. The values in column 3 are always much higher than the corresponding values in columns 1 and 2, which shows that the two lists hold largely different sets of ambiguous words/phrases.
- Column 4 should be viewed as an upper bound on ambiguity found in the logs as the union of all lists of potentially ambiguous words was used including (Wi_D) for which a commonly agreed dominant referent is known.

Name		1 Wi_{ND}	2 WN	3 WN+Wi_{ND}	4 WN+Wi_D+Wi_{ND}
Web	freq	7.6%	4.0%	10.0%	16.4%
	all	2.5%	0.8%	3.0%	3.9%
PA	freq	10.5%	6.4%	14.7%	23.6%
	all	2.1%	0.8%	2.7%	3.7%

Table 6. Fraction of queries matching in combinations of one of three lists

Looking across all columns, it is clear that finding whole queries that are ambiguous is a relatively common event particularly in the more frequent queries submitted to each engine. By including the ambiguous words/phrases from Wikipedia, many more ambiguous queries (particularly when measured across the *all* logs sets) were identified than through the more conventional approach of just using WordNet. Although the percentage figures for ambiguous queries in the *all* set appear small, it should be remembered that the logs are large: e.g. a 3% overlap on the Web log represents 30,000 distinct ambiguous queries.

5.2 Length of ambiguous queries

The next analysis was to examine the average length of the matching ambiguous queries. Defining a space as a word separator, the average number of words in the ambiguous queries was calculated for matches in the *WN* and Wi_{ND} lists. The results are shown in Table 7.

Name		1 Wi_{ND}	2 >1 word	3 WN
Web	freq.	1.13	*10.4%*	1.01
	all	1.26	*20.4%*	1.04
PA	freq.	1.12	*10.8%*	1.02
	all	1.21	*16.3%*	1.03

Table 7. Word length of ambiguous queries

Compared to the average word lengths in Table 3, the averages here are shorter. With most ambiguous queries being one word long, this being particularly true of the most frequent queries in the logs. It is worth noting, however that for those queries matched in the Wikipedia list (column 1), more multi-word ambiguous queries were found. Column 2 shows the percentage of the queries longer than one word.

5.3 Discussion

The results shown here indicate that through analysis of a new source of ambiguous words/phrases, ambiguity in whole queries is common. The ambiguity is present not just for single word queries, but for some multi-word queries as well.

That the results were consistent across the logs of two different search engines strengthens our conclusions. However, it should be remembered that the collections of both engines are large and heterogeneous, which increases the likelihood of finding matches to more than one interpretation of an ambiguous query. Smaller more focused collections, such as a small digital library, might have queries submitted that, according to Wikipedia or WordNet, are ambiguous, but only one interpretation of the query is actually present in the collection, meaning the ambiguity can be ignored.

This study has not examined the actual impact of this ambiguity on retrieval effectiveness, which has to be left for future work. Nevertheless, there are a great many large collections and for search engines retrieving from such corpora, it would appear important for them to deal effectively with ambiguous queries.

6. Testing search engines with ambiguity

It was decided next to conduct an initial exploration of how well current IR systems retrieve documents from queries that are ambiguous. Given an ambiguous query, will a conventional search engine retrieve documents in the top ten ranks that are relevant to more than one of the query's interpretations?

As established in Sec.2.3, there are currently no publicly available test collections with ambiguous queries. Ultimately, we believe it will be necessary to build such a collection. Before engaging in the work of creating such a corpus, it was decided to explore methods of simulating such a collection. To achieve this, an old technique of simulating ambiguity was re-examined.

6.1 Pseudo queries

Sanderson [15] used pseudo-words to explore the impact of ambiguity on IR. A pseudo-word is an artificial term created out of the concatenation of two or more unrelated words. All occurrences of the words in a test collection are replaced with a token that represents the pseudo-word (e.g. replacing "banana" and "magazine" with "banana#magazine"). By introducing the pseudo-word, the collection becomes additionally ambiguous and by varying the number and size of pseudo-words, Sanderson found relationships between ambiguity, disambiguation and retrieval effectiveness.

In order to evaluate conventional search engines retrieving ambiguous queries, it was decided to adapt pseudo-words to create so called *pseudo-queries*: where two or more topics were merged into a single topic. Taking such an approach, it was possible to merge the topics of an existing test collection to form a collection with ambiguous topics that have distinct sets of relevance judgments, one set for each topic interpretation. This was the methodology adopted here.

6.1.1 Are pseudo-queries a good simulation?

At the time of conducting his original work, Sanderson had to defend the use of pseudo-words against the concern that they were a poor simulation of ambiguity. It was argued that treating randomly selected words as legitimate *pseudo-senses* of a new ambiguous word was incorrect as the senses of actual ambiguous words were generally semantically related in some manner. Sanderson [16] justified his methodology by citing Gale et al [10] who showed that typically, the real senses of words occur in different discourses. Sanderson argued that this quality of senses was mimicked by pseudo-words. He also presented experimental results showing strong similarities between certain properties of pseudo-senses and real senses. He concluded that for the purposes of retrieval experiments, pseudo-words were a reasonable simulation of ambiguous words.

The question of the relatedness of sense is less of a concern for the work described here, as the referents of proper nouns are commonly unrelated. Consequently, it was concluded that for the purposes of this work, pseudo-queries provided an effective simulation of ambiguous queries.

6.2 Forming a collection of pseudo queries

The pseudo-query collection was built as follows. The Financial Times (FT) newspaper articles of the TREC collection were indexed. The 150 topics 301-450, which have qrels from the FT, were extracted. A number of topics in TREC are "hard": search engines fail to retrieve many relevant documents for those topics. If a pseudo-query was formed from the pairing of a hard topic with an "easier" topic, it would be challenging for a search engine to ensure that relevant documents were retrieved for both pseudo-senses of the query. For this initial experiment, therefore, it was decided to use a subset of easier topics, which were selected as follows. Using titles only, the 150 topics were run on a search engine (Lemur v4.5) and those for which P@10=0 (precision at 10) were removed. This process left 74 topics whose titles were used to form pseudo-queries.

The merging of two topics was achieved using the "SYN" operator available in Lemur's structured query language. As described in the Lemur manual pages[4] "*The terms of the operator are treated as instances of the same term*", this is equivalent to replacement of words with a pseudo-word. To illustrate, the single word titles of the topics 364 and 349 would merged to form the Lemur structured topic

```
#SYN(rabies Metabolism);
```

To merge multi-word topics, only topics containing the same number of words were paired. To illustrate, topics 340 and 313 would be merged as follows

```
#SUM(#SYN(Viral     Industrial)
     #SYN(Hepatitis Espionage));
```

The components of a "SUM" operator (in this case 2 "SYN" operators) are treated equally by Lemur when it calculates a document ranking. The resulting query simulated a situation where a single topic has one of two possible interpretations: one related to "viral hepatitis" the other to "industrial espionage".

The selection of which topics, with equivalent word length, to pair was chosen at random. Note, because of the restriction of only pairing topics with the same title word count, only 35 pairings were possible from the 74 topics.

6.3 Is this a fair test for Lemur?

One might argue that testing an IR system on a type of query it was never designed to work with is an unfair experiment. However, the test is less of the system and more of the existing test collections the IR community uses. Like most IR systems, Lemur's default ranking algorithm was built to ensure it produces good results on TREC collections. There is a common assumption the topics of such collections are in some way a representative sample of the topics a search engine will operationally encounter. We have shown that ambiguous topics are relatively common, if the topics of a test collection like TREC are representative, Lemur should be able to retrieve ambiguous topics successfully.

6.4 Results

As stated above, the 74 topics selected were those which Lemur retrieved at least one relevant document in the top 10. One might view therefore, the 74 as relatively easy topics: P@10 measured across them unpaired was 0.37.

When the 35 paired ambiguous topics were run on Lemur, the measure of effectiveness used to test the search engine was to determine if at least one relevant document was retrieved in the top 10 for each of the 2 interpretations of the ambiguous topic. As topic pairing was controlled through random selection, the whole experiment was repeated 50 times.

The results were that on average for 25 of the 35 pseudo-queries, no relevant document in the top 10 was retrieved for one of the constituent topics. For 72% of the pseudo-queries, only relevant documents for one of the topic's interpretations were retrieved. Given that each constituent topic was selected because it was relatively easy to retrieve a relevant document in the top 10, this result came as a bit of a surprise. Although it should be remembered experimental search engines such as Lemur have been tuned over many years to perform well on test collections with one interpretation per topic. With that thought in mind, a common technique used by search engines was tested on the newly ambiguous test collection.

6.4.1 Pseudo-relevance feedback

Pseudo-relevance feedback (PRF) is a well known technique: given a query, the method conducts a search, extracts terms from the resulting top ranked documents, expands the original query with the extracted terms and conducts another search, which on average produces better retrieval effectiveness than the initial search. Lemur provides support for PRF and so another experiment using PRF was conducted with same 50 sets of random topic pairings used in the experiment above.

The results of this experiment were that on average for 34 of the 35 topics no relevant document in the top 10 was retrieved for one of the constituent topics. For 97% of the pseudo-queries, retrieval failed for one interpretation. PRF appears drive a retrieval system to one interpretation per ranking and in the context of these queries, this is completely the wrong thing to do.

7. Conclusions

This study examined ambiguity in words and phrases not normally found in dictionaries or thesauri. It was shown that ambiguity is common and that the referents of such words/phrases

[4] www.lemurproject.org/lemur/StructuredQuery.php (Jan. '08)

are often numerous and cover a wide range of types. Such terms were found to be common in the query logs of two different search engines[5]. We conclude from this set of analyses that query ambiguity is a potential problem in many retrieval situations.

A methodology for simulating ambiguous topics was described and a test collection was built. The collection was used to show that a well established experimental IR system does not deal effectively with ambiguous queries. In addition when a process (PRF) normally thought to improve retrieval effectiveness was applied, effectiveness was instead substantially reduced.

Test collections are catalysts for research. As described in Sec. 2.3, the relevance judgments of almost all test collections are based on one person's interpretation of a topic. It would appear this is because almost all collections base their design on the Cranfield model. It was established, however, that it is not unusual for a query to have more than one interpretation per topic and that it is important for search engines to retrieve documents covering each interpretation.

There is a long history of research into methods that address either sub-topics or ambiguous queries (e.g. clustering, maximal marginal relevance, sub-topic retrieval, divergence in ranks, etc.). There is a danger, however, that the true worth of these methods has not been fully realized by the research community because there are no publicly available test collections that have ambiguous topics and a range of relevance judgments that cover more than one interpretation of such topics.

From our study we conclude that new test collections are needed to catalyze research into a generally overlooked though important type of query.

8. ACKNOWLEDGMENTS

Thanks to Paul Clough & Nick Craswell for valuable conversations and data. Funding was provided by the TrebleCLEF project: EU grant number 215231.

9. REFERENCES

[1] Allan, J., Carterette, B., Aslam, J., Pavlu, V., Dachev, B., Kanoulas, E. (2007) Million Query Track 2007 Overview, in *TREC 2007 Notebook*

[2] Artiles, J., Gonzalo, J., Sekine, S. (2007) The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task, *In Proc. of 4th Semeval Workshop*

[3] Carbonell, J.G., Goldstein, J. (1998) The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *Proc SIGIR*, 335-336

[4] Chen, H., Dumais, S (2000) Bringing Order to the Web: Automatically Categorizing Search Results, in *Proc CHI*, 145-152

[5] Subsequent to the completion of this work, we became aware that two additional tags were used to identify disambiguation pages in Wikipedia. The tags identified approximately 20,000 additional ambiguous words (on top of the 89,109 forms in this paper). Not including these words alters the conclusions to the extent that ambiguity in query logs is more common than we calculated and therefore the need for test collections to address such queries is far greater.

[5] Chen, H., Karger, D.R. (2006) Less is more: probabilistic models for retrieving fewer relevant documents, in *Proc SIGIR*, 429-436

[6] Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher S., MacKinnon, I. (2008) Novelty and Diversity in Information Retrieval Evaluation, in *Proc SIGIR*

[7] Cleverdon, C.W. (1991) The Significance of the Cranfield Tests on Index Languages, in *Proc SIGIR*, 3-12

[8] Darwish, K. and Oard, D. (2003) Probabilistic structured query methods, in *Proc SIGIR*, 338-344

[9] Gale W., Church K.W., Yarowsky D. (1992a) Estimating upper and lower bounds on the performance of word-sense disambiguation programs, in *Proc. ACL*, 249-256.

[10] Gale W., Church K.W., Yarowsky D. (1992b) One sense per discourse, in *Proc. of Wrkshp on Speech and Natural Language (HLT Conf.)*, 233-237

[11] Gonzalo, J., Verdejo, F., Chugur, I. and Cigarran, J. (1998) Indexing with WordNet synsets can improve Text Retrieval, *Proc. COLING/ACL Wrkshp on Usage of WordNet for NLP*

[12] Jansen, B.J., Spink, A. (2006) How are we searching the World Wide Web? A comparison of nine search engine transaction logs, *IP&M*, 42(1), 248-263.

[13] Krovetz, R. & Croft, W.B. (1992). Lexical Ambiguity and Information Retrieval, in *ACM TOIS*, 10(1)

[14] Leidner, J. (2007) Toponym Resolution in Text: Annotation, Evaluation & Applications of Spatial Grounding of Place Names. *Ph.D. thesis, School of Informatics, U. Edinburgh*

[15] Sanderson M. (1994) Word sense disambiguation and information retrieval, in Proc. SIGIR, 142-151.

[16] Sanderson, M. & Van Rijsbergen, C.J. (1999) The impact on retrieval effectiveness of skewed frequency distributions, in *ACM TOIS* 17(4), 440-465.

[17] Schütze, H. (1992) Context Space, *In AAAI Fall Symp. on Probabilistic Approaches to Natural Language*, 113-120.

[18] Spärck-Jones, K., Robertson, S.E., Sanderson, M. (2007) Ambiguous requests: implications for retrieval tests, systems and theories, in *ACM SIGIR Forum*

[19] Stokoe, C., Oakes, M.P., Tait J. (2003) Word sense disambiguation in information retrieval revisited. in *Proc. SIGIR*, 159-166

[20] Voorhees, E.M. (1993). Using WordNet™ to disambiguate word sense for text retrieval, in *Proc. SIGIR*, 171-180.

[21] Weiss, S.F. (1973). Learning to disambiguate, in *Information Storage and Retrieval*, 9: 33-41

[22] Willett, P. (1988) Recent trends in hierarchic document clustering: a critical review, *IP&M*, 24(5), 577-597

[23] Zhai, C, Cohen, W.W., Lafferty, J.D. (2003) Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, in *Proc. SIGIR*, 10-17

[24] Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., Ma, W.Y. (2005) Improving web search results using affinity graph. in *Proc. SIGIR*, 504-511.