The Linguistic Structure of English Web-Search Queries

Cory Barr and Rosie Jones
Yahoo! Inc.
701 First Ave
Sunnyvale, CA 94089
`barrc,jonesr@yahoo-inc.com`

Moira Regelson
Perfect Market, Inc.
Pasadena, CA 91103
`mregelson@perfectmarket.com`

Abstract

Web-search queries are known to be short, but little else is known about their structure. In this paper we investigate the applicability of part-of-speech tagging to typical English-language web search-engine queries and the potential value of these tags for improving search results. We begin by identifying a set of part-of-speech tags suitable for search queries and quantifying their occurrence. We find that proper-nouns constitute 40% of query terms, and proper nouns and nouns together constitute over 70% of query terms. We also show that the majority of queries are noun-phrases, not unstructured collections of terms. We then use a set of queries manually labeled with these tags to train a Brill tagger and evaluate its performance. In addition, we investigate classification of search queries into grammatical classes based on the syntax of part-of-speech tag sequences. We also conduct preliminary investigative experiments into the practical applicability of leveraging query-trained part-of-speech taggers for information-retrieval tasks. In particular, we show that part-of-speech information can be a significant feature in machine-learned search-result relevance. These experiments also include the potential use of the tagger in selecting words for omission or substitution in query reformulation, actions which can improve recall. We conclude that training a part-of-speech tagger on labeled corpora of queries significantly outperforms taggers based on traditional corpora, and leveraging the unique linguistic structure of web-search queries can improve search experience.

1 Introduction

Web-search queries are widely acknowledged to be short (2.8 words (Spink et al., 2002)) and to be frequently reformulated, but little else is understood about their grammatical structure. Since search queries are a fundamental part of the information retrieval task, it is essential that we interpret them correctly. However, the variable forms queries take complicate interpretation significantly. We hypothesize that elucidating the grammatical structure of search queries would be highly beneficial for the associated information retrieval task.

Previous work with queries (Allan and Raghavan, 2002) considered that short queries may be ambiguous in their part of speech and that different documents are relevant depending on how this ambiguity is resolved. For example, the word "boat" in a query may be intended as subject of a verb, object of a verb, or as a verb, with each case reflecting a distinct intent. To distinguish between the possibilities, Allan and Raghavan (Allan and Raghavan, 2002) propose eliciting feedback from the user by showing them possible contexts for the query terms.

In addition to disambiguating query terms for retrieval of suitable documents, part-of-speech tagging can help increase recall by facilitating query reformulation. Zukerman and Raskutti (Zukerman and Raskutti, 2002) part-of-speech tag well-formed questions, and use the part-of-speech tags to substitute synonyms for the content words.

Several authors have leveraged part-of-speech tagging towards improved index construction for information retrieval through part-of-speech-

based weighting schemas and stopword detection (Crestani et al., 1998), (Chowdhury and McCabe, 2000), (Dincer and Karaoglan, 2004). Their experiments show degrees of success. Recently, along with term weighting, Lioma has been using part-of-speech n-grams for noise and content detection in indexes (Lioma, 2008). Our study differs from these in that linguistic and part-of-speech focus is almost exclusively placed on queries as opposed to the indexed documents, reflecting our opinion that queries exhibit their own partially predictable and unique linguistic structure different from that of the natural language of indexed documents. Similarly, (Strzalkowski et al., 1998) added a layer of natural language processing using part-of-speech tags and syntactical parsing to the common statistical information-retrieval framework, much like experiments detailed in sections 4 and 5. Our system differs in that our syntactic parsing system was applied to web-search queries and uses rules derived from the observed linguistic structure of queries as opposed to natural-language corpora. By focusing on the part-of-speech distribution and syntactic structure of queries over tagged indexed documents, with a simple bijection mapping our query tags to other tag sets, our system offers a complementary approach that can be used in tandem with the techniques referenced above.

Lima and Pederson (de Lima and Pederson, 1999) conducted related work in which part-of-speech tagging using morphological analysis was used as a preprocessing step for labeling tokens of web-search queries before being parse by a probabilistic context-free grammar tuned to query syntax. We believe this technique and others relying on part-of-speech tagging of queries could benefit from using a query-trained tagger prior to deeper linguistic analysis.

Pasca (Pasca, 2007) showed that queries can be used as a linguistic resource for discovering named entities. In this paper we show that the majority of query terms are proper nouns, and the majority of queries are noun-phrases, which may explain the success of this data source for named-entity discovery.

In this work, we use metrics that assume a unique correct part-of-speech tagging for each query, implicitly addressing the disambiguation issue through inter-annotator-agreement scores and tagger generalization error. To identify these tags, we first analyze the different general forms of queries. In Section 2 we determine a suitable set of part-of-speech labels for use with search queries. We then use manually labeled query data to train a tagger and evaluate its performance relative to one trained on the Brown corpus in Section 3. We make observations about the syntactic structure of web-search queries in Section 4, showing that the majority (70%) of queries are noun-phrases, in contrast with the commonly held belief that queries consist of unstructured collections of terms. Finally, we examine the potential use of tagging in the tasks of search relevance evaluation and query reformulation in Section 5.

2 Data

We sampled queries from the Yahoo! search engine recorded in August 2006. Queries were systematically lower-cased and white-spaced normalized. We removed any query containing a non-ASCII character. Queries were then passed through a high-precision proprietary query spelling corrector, followed by the Penn Treebank tokenizer. No other normalization was carried out. Despite Penn-tokenization, queries were typical in their average length (Jansen et al., 2000). We sampled 3,283 queries from our dataset to label, for a total of 2,508 unique queries comprised of 8,423 individual tokens.

2.1 Inter-rater Agreement

The sparse textual information in search queries presents difficulties beyond standard corpora, not only for part-of-speech tagging software but also for human labelers. To quantify the level of these difficulties we measured inter-rater agreement on a set of 100 queries labeled by each editor. Since one labeler annotated 84.4% of the queries, we used a non-standard metric to determine agreement. One hundred queries were selected at random from each of our secondary labelers. Our primary labeler then re-labeled these queries. Accuracy was then calculated as a weighted average, specifically the mean of the agreement between our primary labeler and secondary labelers, weighted by the number of queries

contributed by each secondary labeler. Measuring agreement with respect to the individual part-of-speech tag for each token, our corpus has an inter-rater agreement of 79.3%. If we require agreement between all tokens in a query, agreement falls to 65.4%. Using Cohen's kappa coefficient, we have that token-level agreement is a somewhat low 0.714 and query-level agreement is an even lower 0.641.

We attempted to accurately quantify token-level ambiguity in queries by examining queries where chosen labels differ. An author-labeler examined conflicting labels and made a decision whether the difference was due to error or genuine ambiguity. Error can be a result of accidentally selecting the wrong label, linguistic misunderstanding (e.g., "chatting" labeled as a verb or gerund), or lack of consensus between editors (e.g., model numbers could be nouns, proper nouns, or even numbers). Examples of genuinely ambiguous queries include "download" and "rent," both of which could be a noun or verb. Another major source of genuine token-level ambiguity comes from strings of proper nouns. For example, some editors considered "stillwater chamber of commerce" one entity and hence four proper-noun tokens while others considered only the first token a proper noun. Of the 99 conflicting token labels in our queries used to measure inter-annotator agreement, 69 were judged due to genuine ambiguity. This left us with a metric indicating query ambiguity accounts for 69.7% of labeling error.

2.2 Tags for Part-of-Speech Tagging Queries

In preliminary labeling experiments we found many standard part-of-speech tags to be extremely rare in web-search queries. Adding them to the set of possible tags made labeling more difficult without adding any necessary resolution. In Table 1 we give the set of tags we used for labeling. In general, part-of-speech tags are defined according to the distributional behavior of the corresponding parts of speech.

Our tag set differs dramatically from the Brown or Penn tag sets. Perhaps most noticeably, the sizes of the tag sets are radically different. The Brown tag set contains roughly 90 tags. In addition, several tags can be appended with additional symbols to indicate negation, genitives, etc. Our tag set contains just 19 unique classes.

Tag	Example	Count (%)
proper-noun	texas	3384 (40.2%)
noun	pictures	2601 (30.9%)
adjective	big	599 (7.1%)
URI	ebay.com	495 (5.9%)
preposition	in	310 (3.7%)
unknown	y	208 (2.5%)
verb	get	198 (2.4%)
other	conference06-07	174 (2.1%)
comma	,	72 (0.9%)
gerund	running	69 (0.8%)
number	473	67 (0.8%)
conjunction	and	65 (0.8%)
determiner	the	56 (0.7%)
pronoun	she	53 (0.6%)
adverb	quickly	28 (0.3%)
possessive	's	19 (0.2%)
symbol	(	18 (0.2%)
sentence-ender	?	5 (0.1%)
not	n't	2 (0.0%)

Table 1: Tags used for labeling part-of-speech in web-search queries.

Our contrasting tag sets reflect an extremely different use of the English language and corresponding part-of-speech distribution. For example, the Brown tag set contains unique tags for 35 types of verbs. We use a single label to indicate all cases of verbs. However, the corpora the Brown tag set was designed for consists primarily of complete, natural-language sentences. Essentially, every sentence contains at least one verb. In contrast, a verb of any type accounts for only 2.35% of our tags. Similarly, the Brown corpus contains labels for 15 types of determiners. This class makes up just 0.66% of our data.

Our most common tag is the proper noun, which constitutes 40% of all query terms, and proper nouns and nouns together constitute 71% of query terms. In the Brown corpus, by contrast, the most common tag, noun, constitutes about 13% of terms. Thus the distribution of tag types in queries is quite different from typical edited and published texts, and in particular, proper nouns are more common than regular nouns.

2.3 Capitalization in Query Data

Although we have chosen to work with lowercase data, web search queries sometimes contain capi-

Use of Capitals	Count	%	Example
Proper-nouns capitalized	48	47%	list of Filipino riddles
Query-Initial-Caps	10	10%	Nautical map
Init-Caps + Proper-Nouns	7	7%	Condos in Yonkers
Acronym	4	4%	location by IP address
Total standard capitalization	69	67%	
All-caps	26	25%	FAX NUMBER FOR ALLEN CANNING CO
Each word capitalized	6	6%	Direct Selling
Mixed	2	2%	SONGS OF MEDONA music feature:audio
Total non-standard capitalization	34	33%	

Table 2: Ways capitalization is used in web-search queries.

talization information. Since capitalization is frequently used in other corpora to identify proper nouns, we reviewed its use in web-search queries. We found that the use of capitalization is inconsistent. On a sample of 290,122 queries from August 2006 only 16.8% contained some capitalization, with 3.9% of these all-caps. To review the use of capitalization, we hand-labeled 103 queries containing capital letters (Table 2).

Neither all-lowercase (83.2%) nor all-caps (3.9%) queries can provide us with any part-of-speech clues. But we would like to understand the use of capitalization in queries with varied case. In particular, how frequently does first-letter capitalization indicate a proper noun? We manually part-of-speech tagged 75 mixed-case queries, which contained 289 tokens, 148 of which were proper nouns. The baseline fraction of proper nouns in this sample is thus 51% (higher than the overall background of 40.2%). A total of 176 tokens were capitalized, 125 of them proper nouns. Proper nouns thus made up 73.3% of capitalized tokens, which is larger than the background occurrence of proper nouns. We can conclude from this that capitalization in a mixed-case query is a fair indicator that a word is a proper noun. However, the great majority of queries contain no informative capitalization, so the great majority of proper nouns in search queries must be uncapitalized. We cannot, therefore, rely on capitalization to identify proper nouns.

With this knowledge of the infrequent use of capital letters in search queries in mind, we will examine the effects of ignoring or using a query's capitalization for part-of-speech tagging in Section 3.4.2.

3 Tagger Accuracy on Search Queries

To investigate automation of the tagging process, we trained taggers on our manually labeled query set. We used 10-fold cross-validation, with 90% of the data used for training and the remaining data used for testing. In the sections below, we used two datasets. The first consists of 1602 manually labeled queries. For the experiments in Section 3.5 we labeled additional queries, for a total of 2503 manually labeled queries.

3.1 Part-of-Speech Tagging Software

We experimented with two freely available part-of-speech taggers: The Brill Tagger (Brill, 1995) and The Stanford Tagger (Toutanova and Manning, 2000; Toutanova et al., 2003).

The Brill tagger works in two stages. The initial tagger queries a lexicon and labels each token with its most common part-of-speech tag. If the token is not in the lexicon, it labels the token with a default tag, which was "proper noun" in our case. In the second stage, the tagger applies a set of lexical rules which examine prefixes, suffixes, and infixes. The tagger may then exchange the default tag based on lexical characteristics common to particular parts of speech. After application of lexical rules, a set of contextual rules analyze surrounding tokens and their parts of speech, altering tags accordingly.

We chose to experiment primarily with the Brill tagger because of its popularity, the human-readable rules it generates, and its easily modifiable code base. In addition, the clearly defined stages and incorporation of the lexicon provide an accessible way to supply external lexicons or entity-detection routines, which could compensate for the sparse contextual information of search queries.

We also experimented with the Stanford Log-Linear Part-of-Speech Tagger, which presently holds the best published performance in the field at 96.86% on the Penn Treebank corpus. It achieves this accuracy by expanding information sources for tagging. In particular, it provides "(i) more extensive treatment of capitalization for unknown words; (ii) features for the disambiguation of the tense forms of verbs; (iii) features for disambiguating particles from prepositions and adverbs." It uses a maximum-entropy approach to handle information

diversity without assuming predictor independence (Toutanova and Manning, 2000).

3.2 Baseline: Most Common Tag

With proper nouns dominating the distribution, we first considered using the accuracy of labeling all tokens "proper noun" as a baseline. In this case, we labeled 1953 of 4759 (41.0%) tokens correctly. This is a significant improvement over the accuracy of tagging all words as "noun" on the Brown corpus (approximately 13%), reflecting the frequent occurrence of proper nouns in search queries. However, to examine the grammatical structure of search queries we must demonstrate that they are not simply collections of words. With this in mind, we chose instead to use the most common part-of-speech tag for a word as a baseline. We evaluated the baseline performance on our manually labeled dataset, with URLs removed. Each token in the set was assigned its most common part of speech, according to the Brill lexicon. In this case, 4845 of 7406 tokens were tagged correctly (65.42%).

3.3 Effect of Type of Training Data

The Brill tagger software is pre-trained on the standard Wall Street Journal corpus, so the simplest possible approach is to apply it directly to the query data set. We evaluated this "out-of-the-box" performance on our 1602 manually labeled queries, after mapping tags to our reduced tag set. (Our effective training-set size is 1440 queries, since 10% were held out to measure accuracy through cross validation.) The WSJ-trained tagger labeled 2293 of 4759 (48.2%) tags correctly, a number well below the baseline performance, demonstrating that application of the contextual rules that Brill learns from the syntax of natural-language corpora has a negative effect on accuracy in the context of queries. When we re-trained Brill's tagger on a manually labeled set of queries, we saw accuracy increase to 69.7%. The data used to train the tagger therefore has a significant effect on its accuracy (Table 3). The accuracy of the tagger trained on query data is above the baseline, indicating that search queries are somewhat more than collections of words.

3.4 Improving Tagger Accuracy

We conducted several experiments in improving tagger accuracy, summarized in Table 3 and described in detail below.

3.4.1 Adding External Lexicon

With a training-set size of 1500 queries, comprising a lexicon of roughly 4500 words, it is natural to question if expanding the lexicon by incorporating external sources boosts performance. To this end, we lower-cased the lexicon of 93,696 words provided by the Brill tagger, mapped the tags to our own tag set, and merged our lexicon from queries. This experiment resulted in an accuracy of 71.1%, a 1.4% increase.

One explanation for the limited increase is that this lexicon is derived from the Brown corpus and the Penn Treebank tagging of the Wall Street Journal. These corpora are based on works published in 1961 and 1989-1992 respectively. As shown in Table 1, proper nouns dominate the distribution of search-engine queries. Many of these queries will involve recent products, celebrities, and other time-sensitive proper nouns. We speculate that Web-based information resources could be leveraged to expand the lexicon of timely proper nouns, thereby enhancing performance.

3.4.2 Experiments with Perfect Capitalization

The overall performance of the pre-trained Brill tagger on our query set may be due to its poor performance on proper nouns, our most frequent part of speech. In the WSJ newspaper training data, proper-nouns always start with a capital letter. As discussed in Section 2.3, capitalization is rare in web-search queries. To examine the effect of the missing capitalization of proper nouns, we evaluated a pre-trained Brill tagger on our previously mentioned manually labeled corpus of 1602 queries altered such that only the proper nouns were capitalized. In this case, the tagger reached an extraordinary 89.4% accuracy (Table 3). Unfortunately, the vast majority of queries do not contain capitalization information and those that do often contain misleading information. The pre-trained tagger achieved only a 45.6% accuracy on non-lowercased queries, performing even worse than on the set with no capitalization at all.

Experiment	Accuracy
Label-all-proper-noun	41.0%
WSJ-trained	48.2%
most-freq-tag-WSJ	64.4%
re-trained	69.7%
retrained + WSJ lexicon	71.1%
user capitalization	45.6%
oracle capitalization	89.4%
automatic capitalization	70.9%

Table 3: Tagging experiments on small labeled corpus. Experiments were conducted on lower-cased queries except where specifically indicated.

3.4.3 Automatic Capitalization

We saw in Section 2.3 that web searchers rarely use capitalization. We have also seen that a pre-trained Brill tagger run on queries with perfect capitalization ("oracle" capitalization) can achieve 89.4% accuracy. We now look at how performance might be affected if we used an imperfect algorithm for capitalization.

In order to attempt to capitalize the proper nouns in queries, we used a machine-learned system which searches for the query terms and examines how often they are capitalized in the search results, weighting each capitalization occurrence by various features (Bartz et al., 2008). Though the capitalization system provides 79.3% accuracy, using this system we see an only a small increase of accuracy in part-of-speech tagging at 70.9%. This system does not improve significantly over the tagger trained on the lower-cased corpus. One explanation is that capitalization information of this type could only be obtained for 81.9% of our queries. Multiplied by accuracy, this implies that roughly 81.9% * 79.3% = 65.0% of our proper nouns are correctly cased. This suggests that any technique for proper-noun detection in search-engine queries must provide over 65.0% accuracy to see any performance increase.

Finally we looked at the capitalization as input by searchers. We trained on the oracle-capitalized corpus, and tested on raw queries without normalization. We saw an accuracy of just 45.6%. Thus using the capitalization input by web searchers is misleading and actually hurts performance.
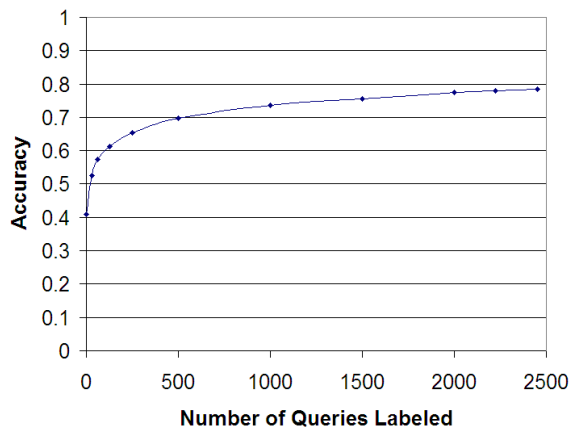

Figure 1: Brill's tagger trained on web-search queries. We see that the most significant gains in performance are with the first few hundred labeled examples, but even after 2500 examples are labeled, more labeled data continues to improve performance.

3.5 Learning Curve

It is important to understand whether tagger accuracy is limited by the small size of our manually labeled dataset. To examine the effect of dataset size, we trained Brill's tagger with increasing numbers of labeled queries and evaluated accuracy with each set size. In the interim between conducting the experiments of sections 3.1 through 3.3 and those of section 3.5, we were able to obtain 1120 new labeled queries, allowing us to extend the learning curve. With our complete corpus of 2722 labeled examples (for a cross-validated training-set size of 2450 labeled examples, URLs omitted), we see an accuracy of 78.6% on a per-token basis. We see the most significant gains in performance with the first few hundred labeled examples, but even after 2500 examples are labeled, more labeled data continues to improve performance.

3.6 Comparing Taggers to Suggest Methods for Boosting Performance

In Table 4 we see a comparison of Brill's tagger to the Stanford tagger trained on 2450 labeled queries. The 0.3% performance increase is not statistically significant. As listed in Section 3, the features the Stanford tagger adds to achieve high accuracy in traditional natural-language corpora are not in-

Tagger	Accuracy
Brill	78.6%
Stanford	78.9%

Table 4: Comparison of Brill's tagger to the Stanford tagger, on our corpus of manually annotated query logs.

formative in the domain of search-engine queries. We believe greater performance on our data will be achieved primarily through examination of common sources of inter-rater disagreement (such as consistent handling of ambiguity) and incorporation of external sources to detect proper nouns not in the lexicon.

To validate our intuition that expanding the lexicon will boost performance, we obtained a proprietary list of 7385 known trademarked terms used in the sponsored-search industry. Treating these phrases as proper nouns and adding them to the lexicon from the Wall Street Journal supplied with the Brill tagger, we see our cross validated accuracy improve to 80.2% (with a standard deviation of 1.85%), the highest score achieved in our experiments. We find it likely that incorporation of more external lexical sources will result in increased performance.

Our experiments also support our hypothesis that addressing inter-annotator agreement will boost performance. We can see this by examining the results of the experiments in section 3.3 verses section 3.5. In section 3.3, we see the accuracy on the query-trained Brill tagger is 69.7%. As mentioned, for the experiment in section 3.5, we were able to obtain 1120 new queries. Each of these newly labeled queries came from the same labeler, who believes their handling of the ambiguities inherent in search queries became more consistent over time. With the same training-set size of 1440 used in section 3.3, Figure 1 shows performance at 1440 queries is roughly 6% higher. We believe this significant improvement is a result of more consistent handling of query ambiguity obtained through labeling experience.

4 Query Grammar

The above-baseline performance of the Brill tagger trained on web-search queries suggests that web-search queries exhibit some degree of syntactical structure. With a corpus of queries labeled with part-of-speech information, we are in a position to analyze this structure and characterize the typical patterns of part-of-speech used by web searchers. To this end, we randomly sampled and manually labeled a set of 222 queries from the part-of-speech dataset used for tagger training mentioned above. Each query was labeled with a single meta-tag indicating query type. Two author-judges simultaneously labeled queries and created the set of meta-tags during much discussion, debate, and linguistic research. A list of our meta-tags and the distribution of each are provided in Table 5. We can see that queries consisting of a noun-phrase dominate the distribution of query types, in contrast with the popularly held belief that queries consist of unstructured collections of terms.

To determine how accurately a meta-tag can be determined based on part-of-speech labels, we created a grammar consisting of a set of rules to rewrite part-of-speech tags into higher-level grammatical structures. These higher-level grammatical structures are then rewritten into one of the seven classes of meta-tags seen in Table 5. Our grammar was constructed by testing the output of our rewrite rules on queries labeled with par-of-speech tags that were not part of the 222 queries sampled for meta-tag labeling. Grammar rules were revised until the failure rate on previously untested part-of-speech-labeled queries stabilized. Failure was evaluated by two means. In the first case, the grammar rules failed to parse the sequence of part-of-speech tags. In the second case, the grammar rules led to an inappropriate classification for a query type. As during the labeling phase, the two author-labelers simultaneously reached a consensus on whether a parse failed or succeeded, rendering an inter-annotator score inapplicable. The resulting grammar was then tested on the 222 queries with query-type meta-tags.

Our rules function much like production rules in context-free grammars. As an example, the two-tag sequence "determiner noun" will be rewritten as "noun phrase." This in turn could be re-written into a larger structure, which will then be rewritten into a meta-tag of query type. The primary difference between a context-free grammar or probabilistic context-free grammar (such as that employed by Lima and Pederson (de Lima and Pederson, 1999))

Query Gramm. Type	Example	Freq (%)
noun-phrase	free mp3s	155 (69.8%)
URI	http:answers.yahoo.com/	24 (10.8%)
word salad	mp3s free	19 (8.1%)
other-query	florida elementary reading conference2006-2007	15 (6.8%)
unknown	nama-nama calon praja ipdn	6 (2.7%)
verb-phrase	download free mp3s	3 (1.4%)
question	where can I download free mp3s	1 (0.45%)

Table 5: Typical grammatical forms of queries used by web searchers, with distribution based on a sample of 222 hand-labeled queries.

and our grammar is that our rules are applied iteratively as opposed to recursively. As such, our grammar yields a single parse for each input.

Some of our rules reflect the telegraphic nature of web queries. For example, it is much more common to see an abbreviated noun-phrase consisting of adjective-noun, than one consisting of determiner-adjective-noun.

Examining the Table 5, we see that just labeling a query "noun-phrase" results in an accuracy of 69.8%. Our grammar boosted this high baseline by 14% to yield an final accuracy result of 83.3% at labeling queries with their correct meta-type. These meta-types could be useful in deciding how to handle a query. Further enhancements to the grammar would likely yield a performance increase. However, we feel accuracy is currently high enough to continue with experiments towards application of leveraging grammar-deduced query types for information retrieval.

We can think of some of these meta-types as elided sentences. For example, the noun-phrase queries could be interpreted as requests of the form "how can I obtain X" or "where can I get information on X", while the verb-phrase queries are requests of the form "I would like to DO-X".

5 Applications of Part-of-Speech Tagging

Since search queries are part of an information retrieval task, we would like to demonstrate that part-of-speech tagging can assist with that task. We conducted two experiments with a large-scale machine-learned web-search ranking system. In addition, we considered the applicability of part-of-speech tags to the question of query reformulation.

5.1 Web Search Ranking

We worked with a proprietary experimental testbed in which features for predicting the relevance of a query to a document can be tested in a machine-learning framework. Features can take a wide variety of forms (boolean, real-valued, relational) and apply to a variety of scopes (the page, the query, or the combination). These features are evaluated against editorial judgements and ranked according to their significance in improving the relevance of results. We evaluated two part-of-speech tag-based features in this testbed.

The first experiment involved a simple query-level feature indicating whether the query contained a noun or a proper noun. This feature was evaluated on thousands of queries for the test. At the conclusion of the test, this feature was found to be in the top 13% of model features, ranked in order of significance. We believe this significance represents the importance of recognizing the presence of a noun in a query and, of course, matching it. Within this experimental testbed a statistically significant improvement of information-retrieval effectiveness is notoriously difficult to attain. We did not see a significant improvement in this metric. However, we feel that our feature's high ranking warrants reporting and hints at a potentially genuine boost in retrieval performance in a system less feature-rich.

The second experiment was more involved and reflected more of our intuition about the likely application of part-of-speech tagging to the improvement of search results. In this experiment, we part-of-speech tagged both queries and documents. Documents were tagged with a conventionally trained Brill tagger with the resulting Penn-style tags mapped to our tag set. Many thousands of query-document pairs were processed in this manner. The feature was based on the percent of times the part-of-speech tag of a word in the query matched the part-of-speech tag of the same word in the document. This feature was ranked in the top 12% by significance, though we again saw no statistically significant increase in overall retrieval performance.

5.2 Query Reformulation

We considered the application of part-of-speech tagging to the problem of query reformulation, in which

Part-of-speech	p(subst)	subst / seen
Number	0.49	148 / 302
Adjective	0.46	2877 / 6299
Noun	0.42	15038 / 35515
Proper noun	0.39	21478 / 55331
Gerund	0.37	112 / 300
Verb	0.31	1769 / 5718
Pronoun	0.23	300 / 1319
Conjunction	0.18	85 / 464
Adverb	0.13	105 / 790
Determiner	0.10	22 / 219
Preposition	0.08	369 / 4574
Possessive	0.08	25 / 330
Not	0.03	1 / 32
Symbol	0.02	16 / 879
Other	0.02	78 / 3294
Sentence-ender	0.01	3 / 234
Comma	0.00	4 / 991

Table 6: Probability of a word being reformulated from one query to the next, by part-of-speech tag. While proper-nouns are the most frequent tag in our corpus, adjectives are more frequently reformulated, reflecting the fact that the proper nouns carry the core meaning of the query.

a single word in the query is altered within the same user session. We used a set of automatically tagged queries to calculate change probabilities of each word by part-of-speech tag and the results are shown in Table 6.

The type of word most likely to be reformulated is "number." Examples included changing a year ("most popular baby names 2007" → "most popular baby names 2008"), while others included model, version and edition numbers ("harry potter 6" → "harry potter 7") most likely indicating that the user is looking at variants on a theme, or correcting their search need. Typically a number is a modifier of the core search meaning. The next most commonly changed type was "adjective," perhaps indicating that adjectives can be used to refine, but not fundamentally alter, the search intent. Nouns and proper nouns are the next most commonly modified types, perhaps reflecting user modification of their search need, refining the types of documents retrieved. Other parts of speech are relatively seldom modified, perhaps indicating that they are not

viewed as having a large impact on the documents retrieved.

We can see from the impact of the search engine ranking features and from the table of query reformulation likelihood that making use of the grammatical structure of search queries can have an impact on result relevance. It can also assist with tasks associated with improving recall, such as query reformulation.

6   Conclusion

We have quantified, through a lexicostatistical analysis, fundamental differences between the natural language used in standard English-language corpora and English search-engine queries. These differences include reduced granularity in part-of-speech classes as well as the dominance of the noun classes in queries at the expense of classes such as verbs frequently found in traditional corpora. In addition, we have demonstrated the poor performance of taggers trained on traditional corpora when applied to search-engine queries, and how this poor performance can be overcome through query-based corpora. We have suggested that greater improvement can be achieved by proper-noun detection through incorporation of external lexicons or entity detection. Finally, in preliminary investigations into applications of our findings, we have shown that query part-of-speech tagging can be used to create significant features for improving the relevance of web search results and may assist with query reformulation. Improvements in accuracy can only increase the value of POS information for these applications. We believe that query grammar can be further exploited to increase query understanding and that this understanding can improve the overall search experience.

References

James Allan and Hema Raghavan. 2002. Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of SIGIR*, pages 307–314.

Kevin Bartz, Cory Barr, and Adil Aijaz. 2008. Natural language generation in sponsored-search advertisements. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 1–9, Chicago, Illinois.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case

study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Abdur Chowdhury and M. Catherine McCabe. 2000. Improving information retrieval systems using part of speech tagging.

Fabio Crestani, Mark Sanderson, and Mounia Lalmas. 1998. Short queries, natural language and spoken document retrieval: Experiments at glasgow university. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, pages 667–686.

Erika F. de Lima and Jan O. Pederson. 1999. Phrase recognition and expansion for short, precision-biased queries based on a query log. In *Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–152, Berkeley, California.

Bekir Taner Dincer and Bahar Karaoglan. 2004. The effect of part-of-speech tagging on ir performance for turkish. pages 771–778.

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227.

Christina Amalia Lioma. 2008. *Part of speech N-grams for information retrieval*. Ph.D. thesis, University of Glasgow, Glasgow, Scotland, UK.

Marius Pasca. 2007. Weakly-supervised discovery of named entities using web search queries. In *CIKM*, pages 683–690.

Amanda Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. 2002. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–109.

Tomek Strzalkowski, Jose Perez Carballo, and Mihnea Marinescu. 1998. Natural language information retrieval: Trec-3 report. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, page 39.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *COLING*, pages 1177–1183, Taipei, Taiwan.