DOES CAPITALIZATION MATTER IN WEB SEARCH?

Silviu Cucerzan
Microsoft Research, 1 Microsoft Way, Redmond, U.S.A

Keywords: Web search, Queries, Capitalization, Truecasing, Ranking.

Abstract: We investigate the capitalization features of queries submitted to Web search engines and the relation between capitalization information, either as received from users or as hypothesized based on Web statistics, and search relevance. We observe that users tend to lowercase words in their queries significantly more often than as predicted from Web data. More importantly, we determine that document relevance is strongly correlated with the matching in capitalization between the instances of query tokens in the target document and the tokens of the truecased form of the query as obtained by using Web n-gram data.

1 INTRODUCTION AND RELATED WORK

Case is an orthographic feature present in Indo-European languages, most of which employ the Latin alphabet. The vast majority of such languages capitalize the first letter of words in proper nouns and the first letter of the first word in a sentence. Additionally, there exist many language dependent and/or stylistic capitalization rules; for example, the names of days are capitalized in English (e.g., “Thursday”), while they are written in lowercase in French (e.g., “jeudi”); words’ capitalization may depend on whether they appear in titles and headings or in running text; etc. In English – the language on which we focus in this study – case information is very useful in disambiguating or reducing the ambiguity of a large number of polysemous words, such as “apple”, “bush”, “turkey”, and “us”, and a very useful feature for several language processing tasks, as shown by Liţă et al. (2003).

Web data, as captured in the Google 1T 5-gram corpus (Brants and Franz, 2006), shows a large variety of capitalizations for almost all English words. More than 81% of the 137,000 words in a large English thesaurus are seen in the Google unigram data set with at least two capitalization forms, as shown in Figure 1. For example, the word “friends” has 31 different capitalizations with at least 200 occurrences: 69 million instances of the form “friends”, 39 million of the form “Friends”, 2 million of “FRIENDS”, and over 35 thousands of various mixed-cased versions. The fact that 9,438 words (e.g., “abandonees” and “carbonizations”) did not appear with any capitalization in the Google unigram set is likely due to the 200 cut-off employed for unigram statistics.

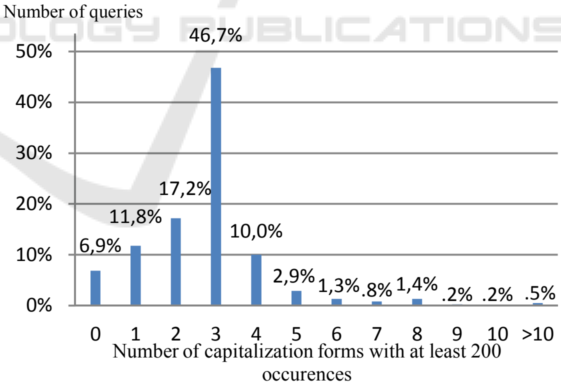

Figure 1: Percentages of words in an English thesaurus seen with various numbers of distinct capitalizations on the Web.

Despite the rich capitalization diversity on the Web, or possibly because of it, all major commercial Web search engines in current use (Google, Yahoo, Bing, and Ask) employ case-insensitive strategies for retrieving Web search results for user queries. This means that differently cased queries, such as “best buy motorcycles” and “Best Buy motorcycles”, which may refer to different topics (best reviewed motorcycles and the retailer’s recent

Cucerzan S..
DOES CAPITALIZATION MATTER IN WEB SEARCH?.
DOI: 10.5220/0003102503020306
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 302-306
ISBN: 978-989-8425-28-7

announcement about selling electric motorcycles, respectively), return the same sets of search results.

Moreover, query logs show that a large percentage of the queries submitted by users contain case information, possibly for multiple reasons: users employ the orthography they typically use in document editing, they perceive the use of uppercase as appropriate when querying for people or geo-political entities, they try to enforce a certain disambiguation of a queried term, they copy and paste substrings from properly-cased documents, or simply by mistake.

Previous research on truecasing has mainly investigated text corpora truecasing, with focus on language processing tasks such as named entity recognition and machine translation (Liță et al., 2003), speech transcription (Chelba and Acero, 2004), sentence boundary detection and casing of words that start sentences (Mikev, 1999), and language dynamics (Batista et. al, 2008). Cucerzan (2010) showed that the capitalization information from Web search snippets can be employed for cross-corpus case normalization. However, the findings of Church (1995) on the effects of text normalization in information retrieval, including case-sensitive search, were inconclusive.

2 DATA COLLECTION

To determine whether capitalization could be an informative ranking feature, we started with a set of 10k distinct queries sampled at random by frequency from the logs of a major search engine, for which relevance judgments on a scale from 0 to 5 were available. We kept for our experiments only the 9,810 queries that include at least one character of the English alphabet (a–z or A–Z). These queries contain a total number of 13,904 word types accounting for 29,388 tokens (average of 3 tokens per query). The longest query has 34 tokens. Most queries have 2 tokens. Figure 2 shows a histogram of the number of tokens per query for the whole set. The most frequent words in the set are “of” and “in”, with 324 and 316 occurrences, respectively, each seen with three capitalization forms in our query set. 21,973 of the query tokens in our set are all lowercase, 3,103 tokens are in mixed case, and 1,794 tokens are in all uppercase, while 477 tokens are numbers. The other 2,041 tokens contain at least one non-English letter or punctuation sign or are a mixture of letters and numbers.

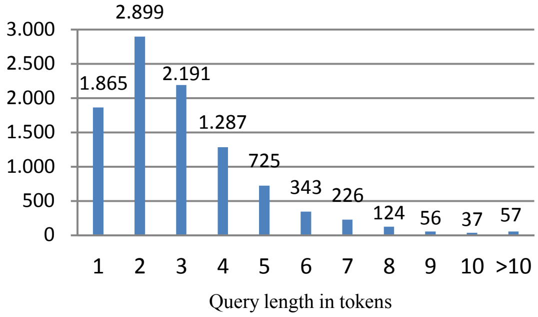

Figure 2: Histogram of query length in tokens.

3 EXPERIMENTS AND FINDINGS

We first investigate how well the capitalization employed by users for the queries in our sample set matches the capitalization statistics of the Web data crawled by Google. To do so, we ignore the tokens that are numbers or contain non-English letters and focus only on the 26,870 query tokens formed exclusively of English letters (referred to as *literal tokens* henceforth). An interesting finding is that only 85.4% of these literal tokens are present in the large English thesaurus employed. The relatively high out-of-vocabulary rate is due to proper nouns (such as “myspace” and “Millau”), foreign words (e.g., “palangoje” and “Konzert”), and misspellings (e.g., “Geroge” and “helecopter”). However, no fewer than 26,311 literal tokens (or 97.9% of the literal tokens in our set) are present with at least one capitalization form in the Google unigram list. On average, Google’s unigram data contains 6.9 distinct capitalizations per literal token from our query set.

For those literal tokens present in the Google unigram set, the users’ capitalization matches the most frequent capitalization form on the Web 47.2% of the time. The matching percentage increases to 54.8% when the literal token contains at least one uppercase letter, but even this number does not seem to indicate reliable signal in the users’ capitalization of queries. Overall, in 84.4% of the capitalization mismatches, the query token was in lowercase, while the most frequent capitalization form had a different capitalization form: upper case in 8% of those instances, the first letter only in uppercase 74.6% of the time, or another mixed case form in 17.4% of the cases. These findings seem to point out that a large number of Web search users tend to write queries in lowercase (thus, matching the case-insensitive models of the Web search engines), which is not surprising. More unexpected is that even when the

Web search engine users employ uppercase letters in their queries, their capitalization matches the most frequent form on the Web rather randomly.

However, since word capitalization is highly dependent on the context in which a word is used, we must also employ higher order n-gram statistics. For each token w_i in a query $q = w_1 \ldots w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2} \ldots w_n$, we examine the statistics for all possible bigrams (left: $w_{i-1} w_i$ and right: $w_i w_{i+1}$) and trigrams (left: $w_{i-2} w_{i-1} w_i$, middle: $w_{i-1} w_i w_{i+1}$, and right: $w_i w_{i+1} w_{i+2}$) that contain it (obviously, some of these are undefined for values of $n \leq 2$ or $i \in \{0, 1, n\text{-}1, n\}$, and thus, cannot be accounted for). 86.8% of the bigrams and 59.8% of the trigrams present in our queries appear in the corresponding Google n-gram sets with at least one capitalization.

To compute the most likely capitalization of a token in a given n-gram based on Google's Web data, we aggregate the Google n-gram counts by folding the case of all other tokens in the n-gram. We find that at bigram level, capitalization of literal tokens in our query set matches the most frequent capitalization in the Google set 53.6% of the time for left bigrams and 55.8% for right bigrams. The matching improves to 59.4% for left trigrams, 64.8% for middles trigrams, and 62.1% for right trigrams. Nonetheless, these numbers are all significantly lower than those obtained by hypothesizing that the capitalization of all tokens is lowercase (mid to high 60s). This indicates that users favour lowercase forms in queries to a higher degree than as predicted by employing Web-based n-gram capitalization statistics.

We now investigate whether capitalization information may be useful for ranking, either as submitted by users or as predicted based on Web n-gram data. For the latter, we employ a system that truecases each query token by using aggregate capitalization counts for all trigrams that contain it, with back-off to the bigrams, and finally to unigrams when higher-order n-grams are undefined or statistics for those n-grams are not available in the Google data. Explicitly, for queries with only one token, we choose the most frequent capitalization of the token in the Google unigram data. For queries with two tokens, the system predicts for each token the most likely capitalization obtained through the process of case folding of the other token and aggregation described above. We back-off to unigram statistics when the bigram does not appear in the Google data set. Similarly, for each token in queries of length 3 or more, the system combines the counts obtained using the case-folding and aggregation process for each possible position of the token in a trigram (left, middle, and right), with back-off to bigrams and unigrams.

Table 1: Capitalization inter-agreement ratios at query level (i.e., the capitalization of all tokens in a query matches).

	Annotator 1	Annotator 2	Original	System
Annotator 1		80%	36%	54%
Annotator 2	80%		33%	48%
Original	36%	33%		28%
System	54%	48%	28%	

Table 2: Capitalization inter-agreement ratios at query token level.

	Annotator 1	Annotator 2	Original	System
Annotator 1		85.5%	61.7%	73.9%
Annotator 2	85.5%		54.5%	70.0%
Original	61.7%	54.5%		49.2%
System	73.9%	70.0%	49.2%	

To estimate how well this truecasing system works, we selected 100 queries at random from our set (Appendix 1), stripped the case information, and asked two annotators to truecase them according to their best guess of the original query intent. Tables 1 and 2 summarize the annotator inter-agreement, as well the matching with the original capitalization and the system-predicted capitalization at query level and token level, respectively. Evidently, percentages are much higher when agreement is computed at token level, as for two queries to match we require that the capitalizations of all component tokens match.

An important observation is that the truecasing system based on the Google n-gram data agrees with the annotators to a much higher degree than its agreement with the original casing of the query, as well as the agreement observed between the annotators' capitalizations and the original capitalization of the queries. We also note that this system predicts a higher number of tokens as starting in uppercase than the human annotators (64.9% and 78.5% of the disagreements with the two annotators at token-level are of this type), which may indicate a Web bias towards capitalized forms.

Finally, we measure the correlation between relevance and the matching of capitalization in queries and documents. For every query and document pair, we compute the percentage of time the capitalization of tokens in the query matches the capitalization forms of the tokens in the text of the document, then we macro-average the obtained values, first at query-document level, and then for all query-document pairs in each of the given relevance categories (from 0 – very bad to 5 – excellent).

Table 3: Matching percentages between capitalization of tokens in queries and documents for each relevance category, as well as the correlation coefficients between the relevance labels/values and these percentages.

Capitalization Type	Relevance Label						Corr.
	0	1	2	3	4	5	
All low	29.0%	28.5%	26.7%	26.3%	20.3%	18.7%	-.94
Original	48.3%	33.6%	33.8%	33.8%	28.0%	23.2%	-.90
Annot. 1	53.9%	60.8%	61.9%	61.8%	61.9%	64.4%	.83
Annot. 2	53.9%	60.5%	61.5%	62.6%	65.5%	62.7%	.82
System	46.8%	66.3%	67.4%	68.1%	71.5%	71.9%	.81

The capitalization hypothesized by the system trained on Google n-gram data matched overall the best the capitalization in the documents in our set, as shown in Table 3. This is not surprising under the assumption that the documents in our set follow the overall capitalization distributions on the Web. However, more importantly and less expected is the very strong positive correlation (0.81) between document relevance labels and capitalization matching for the queries truecased by the system. Similar correlation coefficients (0.82 and 0.83) are also seen when using the annotators' cased versions of the queries. On the opposite, the matching of the original user capitalization is strongly negatively correlated (-0.9) with the relevance values, which may explain at least to some degree why query capitalization is typically perceived as inadequate in Web search ranking (to the best of our knowledge). Moreover, the strong negative correlation (-0.94) between the all-lowercase query form and document relevance provides another empirical confirmation to the fact that documents in which the query words are capitalized (possibly because they are in titles or headings) tend to be more relevant for the target query.

4 DISCUSSION

To determine the concrete impact of using query truecasing and capitalization features in Web search, more costly actual ranking experiments are needed. However, the strong correlation observed in our experiments between the quality of candidate Web documents for a query and the matching of the capitalization of the query tokens in the truecased from of the query and in the candidate documents indicates that capitalization information could be very important for ranking, and warrants such ranking experiments.

While case-sensitive indexing of Web pages would present numerous implementation disadvantages and could also lower recall to a substantial degree, the implementation of a system as suggested in this paper requires only modifying the data structures of the inverted index of the search engine to store for each word instance two additional bits that encode the capitalization of that instance in the indexed document (lower case, all uppercase, first uppercase, or other mixed casing) and would have no impact on recall.

5 CONCLUSIONS

We analyzed the capitalization of a random sample of queries submitted by users of a major commercial Web search engine. As expected, we observed that users tend to lowercase their queries significantly more often than as predicted from Web n-gram data. We also showed that by employing Web n-gram statistics to truecase the user queries, we obtain query forms for which query-document capitalization matching is strongly positively correlated with document relevance for the target query. This result indicates that capitalization features could be employed beneficially in Web search ranking.

REFERENCES

Batista, F., Marmede, N., and Trancoso, I. 2008. Language Dynamics and Capitalization using Maximum Entropy. In *Proceedings of ACL 2008: HLT Companion volume*, pages 1-4.

Brants, T. and Franz, G. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Catalog ID: LDC2006T13.

Chelba, C. and Acero, A. 2004. Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot. In *Proceedings of EMNLP 2004*, pages 285-292.

Church, K. 1995. One Term or Two? In *Proceedings of SIGIR 1995*, pages 310-318.

Cucerzan, S. 2010. A Case Study of Using Web Search Statistics: Case Restoration. In *Proceedings of CICLing 2010*, *LNCS 6008*, pages 199-211.

Liţă, L. V., Ittycheriah, A., Roukos, S., and Kambhatla, N. 2003. tRuEcasIng. In *Proceedings of ACL 2003*, pages 152–159.

Mikev, A. 1999. A Knowledge-free Method for Capitalized Word Disambiguation. In *Proceedings of ACL 1999*, pages 159–166.

APPENDIX

Appendix 1: Random sample of 100 queries on which inter-agreement and matching statistics are reported. The queries are shown with the original capitalization submitted by users. Original spelling was preserved; space and punctuation were normalized. Two queries (marked with ***) were anonymized for privacy reasons by changing the last names originally present in the queries, but preserving the case information.

windows down loads	ronald reagan
University of Miami , Ohio	pregnancy safe hair color
buy bigger house or stay	wayne *doe* ***
chaminade university	john *doe* , England ***
tony stewart wallpaper	furniture auctions
food lion auto fair	time zones in the united states
easports	msn music
handmade paper " elephant "	crossword puzzles
hacienda puerto rico tax return	onofrio dog shows
I485 form	embryonic stem cell & diabetes
j and r	staplescenter
BELGIUM HORSE MILK	FIRST TIME HOME BUYERS PROGRAM MISSISSIPPI
M & T	" India and Africa "
quest medical lab nj	GASTROENTERITIS
what channel is hammy the hamster on in the usa	6 week makeover
cesar quintero	lonestar steakhouse
youth football helments	celeste
childrenplace	esp hobby
idaho springs	iowa sex offender registry
metabolism	online casino with lots of slots free
HP Hard Drive	anxiety overview
free quick budget tools ramsey	" masha kirilenko "
Retroflex lateral flap	1973 4 speed cutlass
usa postal codes	fighter aircraft of vietnam
forty - five seconds inside a tornado	ontario tourism
keystone	FAT ATTACK COMBO
Jenny Saville art	pa lottery
cork strip	HELL ANGELS MC
enolmatic	Powassan virus
laptop memory	Eastbay coupons
whitworth council	Naruto music mp3s
news in south florida	General Bandages
volvo c30	digging for the truth roanoke the lost colony
listen to grillz	judge judy
bank one	power acoustics amplifiers
anchorage hotel ogunquit	google finance
contra costa county animal servies	hip replacement recovery
neo pharm	crate and barrel
tonic you wanted more lyrics	hotwheels
robbins brothers jewerly	wa state national guard
kdka	truck trader on line
Communist diggers find hell	coast guard
Peoria Illinois Doppler Radar	abercrombie
classification of conflicts	fertility calculator
trinity alps	Oscar predictions Foreign Language Film
willard brothers auto sales	Six Flags over Texas
aquarium fish	anime
JPA	dogpile
weatherchannel	halmoon ny sherrif department
aim	stuffing envelopes at home