Crowd Prefers the Middle Path: A New IAA Metric for Crowdsourcing Reveals Turker Biases in Query Segmentation

Rohan Ramanath[*]
R. V. College of Engineering
Bangalore, India
ronramanath@gmail.com

Monojit Choudhury
Microsoft Research Lab India
Bangalore, India
monojitc@microsoft.com

Kalika Bali
Microsoft Research Lab India
Bangalore, India
kalikab@microsoft.com

Rishiraj Saha Roy[†]
Indian Institute of Technology Kharagpur
Kharagpur, India
rishiraj@cse.iitkgp.ernet.in

Abstract

Query segmentation, like text chunking, is the first step towards query understanding. In this study, we explore the effectiveness of crowdsourcing for this task. Through carefully designed control experiments and Inter Annotator Agreement metrics for analysis of experimental data, we show that crowdsourcing may not be a suitable approach for query segmentation because the crowd seems to have a very strong bias towards dividing the query into roughly equal (often only two) parts. Similarly, in the case of hierarchical or nested segmentation, turkers have a strong preference towards balanced binary trees.

1 Introduction

Text chunking of Natural Language (NL) sentences is a well studied problem that is an essential preprocessing step for many NLP applications (Abney, 1991; Abney, 1995). In the context of Web search queries, *query segmentation* is similarly the first step towards analysis and understanding of queries (Hagen et al., 2011). The task in both the cases is to divide the sentence or the query into contiguous *segments* or chunks of words such that the words from a segment are related to each other more strongly than words from different segments (Bendersky et al., 2009). It is typically assumed that the segments are structurally and semantically coherent and, therefore, the information contained in them can be processed holistically.

f	Pipe representation	*Boundary var.*
4	apply | first aid course | on line	1 0 0 1 0
3	apply first aid course | on line	0 0 0 1 0
2	apply first aid | course on line	0 0 1 0 0
1	apply | first aid | course | on line	1 0 1 1 0

Table 1: Example of flat segmentation by Turkers. *f* is the frequency of annotations; segment boundaries are represented by |.

f	Bracket representation	*Boundary var.*
4	((apply first) ((aid course) (on line)))	0 2 0 1 0
2	(((apply (first aid)) course) (on line))	1 0 2 3 0
2	((apply ((first aid) course)) (on line))	2 0 1 3 0
1	(apply (((first aid) course) (on line)))	3 0 1 2 0
1	((apply (first aid)) (course (on line)))	1 0 2 1 0

Table 2: Example of nested segmentation by Turkers. *f* is the frequency of annotations.

A majority of work on query segmentation relies on manually segmented queries by human experts for training and evaluation of segmentation algorithms. These are typically small datasets and even with detailed annotation guidelines and/or close supervision, low Inter Annotator Agreement (IAA) remains an issue. For instance, Table 1 illustrates the variation in flat segmentation by 10 annotators. This confusion is mainly because the definition of a segment in a query is ambiguous and of an unspecified granularity. This is further compounded by the fact that other than easily recognizable and agreed upon segments such as Named Entities or Multi-Word Expressions, there is no established notion of linguistic grouping such as phrases and clauses in a query.

Although there is little work on the use of crowdsourcing for query segmentation (Hagen et al., 2011; Hagen et al., 2012), the idea that the

[*]The work was done during author's internship at Microsoft Research Lab India.

[†]This author was supported by Microsoft Corporation and Microsoft Research India under the Microsoft Research India PhD Fellowship Award.

crowd could be a potential (and cheaper) source for reliable segmentation seems a reasonable assumption. The need for larger datasets makes this an attractive proposition. Also, a larger number of annotations could be appropriately distilled to obtain better quality segmentations.

In this paper we explore crowdsourcing as an option for query segmentation through experiments designed using Amazon Mechanical Turk (AMT)[1]. We compare the results against gold datasets created by trained annotators. We address the issues pertaining to disagreements due to both ambiguity and granularity and attempt to objectively quantify their role in IAA. To this end, we also conduct similar annotation experiments for NL sentences and randomly generated queries. While queries are not as structured as NL sentences they are not simply a set of random words. Thus, it is necessary to compare query segmentation to the über-structure of NL sentences as well as the unter-structure of random *n*-grams. This has important implications for understanding any inherent biases annotators may have as a result of the apparent lack of structure of the queries.

To quantify the effect of granularity on segmentation, we also ask annotators to provide hierarchical or nested segmentations for real and random queries, as well as sentences. Following Abney's (1992) proposal for hierarchical chunking of NL, we ask the annotators to group *exactly two* words or segments at a time to recursively form bigger segments. The concept is illustrated in Fig. 1. Table 2 shows annotations from 10 Turkers. It is important to constrain the joining of exactly two segments or words at a time to avoid the issue of fuzziness in granularity. We shall refer to this style of annotation as *Nested segmentation*, whereas the non-hierarchical non-constrained chunking will be referred to as *Flat segmentation*.

Through statistical analysis of the experimental data we show that crowdsourcing may not be the best practice for query segmentation, not only because of ambiguity and granularity issues, but because there exist very strong biases amongst annotators to divide a query into two roughly equal parts that result in misleadingly high agreements. As a part of our analysis framework, we introduce a new IAA metric for comparison across flat and nested segmentations. This versatile metric can be readily adapted for measuring IAA for other linguistic annotation tasks, especially when done using crowdsourcing.

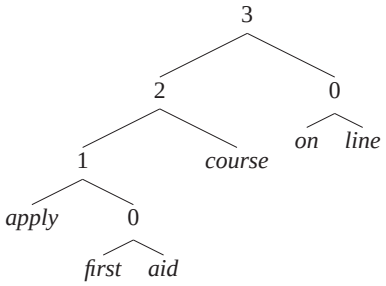

Figure 1: Nested Segmentation: Illustration.

The rest of the paper is organized as follows. Sec 2 provides a brief overview of related work. Sec 3 describes the experiment design and procedure. In Sec 4, we introduce a new metric for IAA, that could be uniformly applied across flat and nested segmentations. Results of the annotation experiments are reported in Sec 5. In Sec 6, we analyze the possible statistical and linguistic biases in annotation. Sec 7 concludes the paper by summarizing the work and discussing future research directions. All the annotated datasets used in this research are freely available for non-commercial research purposes[2].

2 Related Work

Query segmentation was introduced by Risvik et. al. (2003) as a possible means to improve Information Retrieval. Since then there has been a significant amount of research exploring various algorithms for this task and its use in IR (see Hagen et. al. (2011) for a survey). Most of the research and evaluation considers query segmentation as a process analogous to identification of phrases within a query which when put within double-quotes (implying exact matching of the quoted phrase in the document) leads to better IR performance. However, this is a very restricted view of the process and does not take into account the full potential of query segmentation.

A more generic notion of segments leads to diverse and ambiguous definitions, making its evaluation a hard problem (see Saha Roy et. al. (2012) for a discussion on issues with evaluation). Most automatic segmentation techniques (Bergsma and Wang, 2007; Tan and Peng, 2008; Zhang et al.,

[1] `https://www.mturk.com/mturk/welcome`

[2] Related datasets and supplementary material can be accessed from `http://bit.ly/161Gkk9` or can be obtained by directly emailing the authors.

2009; Brenes et al., 2010; Hagen et al., 2011; Li et al., 2011) have so far been evaluated only against a small set of human-annotated queries (Bergsma and Wang, 2007). The reported low IAA for such datasets casts serious doubts on the reliability of annotation and the performance of the algorithms evaluated on them (Hagen et al., 2011; Saha Roy et al., 2012).

There has been a commendable effort by Hagen et al. (2011) to generate a large dataset of $50,000$ (flat) segmented queries (Webis-QSeC-10) obtained through AMT[3]. However, they noticed a lot of spammers and initial disagreements amongst the Turkers and had to spend a lot of time to systematically identify the spammers and eradicate the annotations. This shows that crowdsourced annotations for query segmentation are not readily usable (known through a personal communication with the authors). Nevertheless, if large scale data has to be procured, crowdsourcing seems to be the only efficient and effective model for this task. It has been proven to be so for other IR and linguistic annotations; see Carvalho et al. (2011) for examples of crowdsourcing for IR resources, and Snow et al. (2008) and Callison-Burch (2009) for language resources.

In the context of NL text, segmentation has been traditionally referred to as *chunking* and is a well-studied problem. Abney (1991; 1992; 1995) defines a chunk as a sub-tree within a syntactic phrase structure tree corresponding to Noun, Prepositional, Adjectival, Adverbial and Verb Phrases. Similarly, Bharati et al. (1995) defines it as Noun Group and Verb Group based only on local surface information. However, cognitive and annotation experiments for chunking of English (Abney, 1992) and other language text (Bali et al., 2009) have shown that native speakers agree on major clause and phrase boundaries, but may not do so on more fine-grained chunks. One important implication of this is that annotators are expected to agree more on the higher level boundaries for nested segmentation than the lower ones. We note that hierarchical query segmentation was proposed for the first time by Huang et al. (2010), where the authors recursively split a query (or its fragment) into exactly two parts and evaluate the final output against human annotations.

[3] `http://www.webis.de/research/corpora`

3 Experiments

The annotation experiments have been designed to systematically study the various aspects of query segmentation. In order to verify the effectiveness and reliability of crowdsourcing, we designed an AMT experiment for flat segmentation of Web search queries. As a baseline, we would like to compare these annotations with those from human experts trained for the task. We shall refer to this baseline as the *Gold annotation* set. Since we believe that the issue of granularity could be the prime reason for previously reported low IAA for segmentation, we also designed AMT-based nested segmentation experiments for the same set of queries, and obtained the corresponding gold annotations.

Finally, to estimate the role of ambiguity inherent in the structure of Web search queries on IAA, we conducted two more control experiments, both through crowdsourcing. First, flat and nested segmentation of well-formed English, i.e., NL sentences of similar length distribution; and second, flat and nested segmentation of randomly generated queries. Higher IAA for NL sentences would lead us to conclude that ambiguity and lack of structure in queries is the main reason for low agreements. On the other hand high or comparable IAA for random queries would mean that annotations have strong biases.

Thus, we have the following four pairs of annotation experiments: flat and nested segmentation of queries from crowdsourcing, corresponding flat and nested gold annotations, flat and nested segmentation of English sentences from crowdsourcing, and flat and nested segmentations for randomly generated queries through crowdsourcing.

3.1 Dataset

For our experiments, we need a set of Web search queries and well-formed English sentences. Furthermore, for generating the random queries, we will use search query logs to learn n-gram models. In particular, we use the following datasets:

Q500, QG500: Saha Roy et al. (2012) released a dataset of 500 queries, 5 to 8 words long, for evaluation of various segmentation algorithms. This dataset has flat segmentations from three annotators obtained under controlled experimental settings, and can be considered as *Gold* annotations. Hence, we select this set for our experiments as well. We procured the corresponding nested

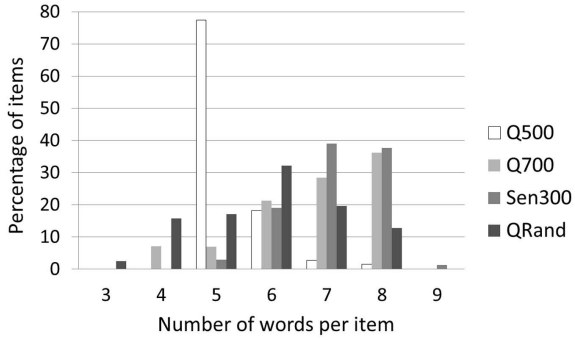

Figure 2: Length distribution of datasets.

segmentation for these queries from two human experts, who are regular search engine users, between 20 and 30 years old, and familiar with various linguistic annotation tasks. They annotated the data under supervision. They were trained and paid for the task. We shall refer to the set of flat and nested gold annotations as **QG500**, whereas **Q500** will be reserved for AMT experiments.

Q700: Since 500 queries may not be enough for reliable conclusion and since the queries may not have been chosen specifically for the purpose of annotation experiments, we expanded the set with another 700 queries sampled from a slice of the query logs of Bing Australia[4] containing 16.7 million queries issued over a period of one month (May 2010). We picked, uniformly at random, queries that are 4 to 8 words long, have only English letters and numerals, and a high *click entropy* because "a query with a larger click entropy value is more likely to be an informational or ambiguous query" (Dou et al., 2008). **Q500** consists of tailish queries with frequency between 5 and 15 that have at least one multiword named entity; but unlike the case of **Q700**, click-entropy was not considered during sampling. As we shall see, this difference is clearly reflected in the results.

S300: We randomly selected 300 English sentences from a collection of full texts of public domain books[5] that were 5 to 15 words long, and checked them for well-formedness. This set will be referred to as **S300**.

QRand: Instead of generating search queries by throwing in words randomly, we thought it will be more interesting to explore annotation of queries generated using n-gram models for $n = 1, 2, 3$. We estimated the models from the Bing Australia log of 16.7 million queries. We generated 250 queries each of desired length distribution using the 1, 2 and 3-gram models. We shall refer to these as **U250**, **B250**, **T250** (for Uni, Bi and Trigram) respectively, and the whole dataset as **QRand**. Fig. 2 shows the query and sentence length distribution for the various sets.

[4] `http://www.bing.com/?cc=au`
[5] `http://www.gutenberg.org`

Parameter	Flat Details	Nested Details
Time needed: actual (allotted)	49 sec (10 min)	1 min 52 sec (15 min)
Reward per HIT	$0.02	$0.06
Instruction video duration	26 sec	1 min 40 sec
Turker qualification	Completion rate >100 tasks	
Turker approval rate	Acceptance rate >60 %	
Turker location	United States of America	

Table 3: Specifics of the HITs for AMT.

3.2 Crowdsourcing Experiments

We used AMT to get our annotations through crowdsourcing. Pilot experiments were carried out to test the instruction set and examples presented. Based on the feedback, the precise instructions for the final experiments were designed.

Two separate AMT Human Intelligence Tasks (HITs) were designed for flat and nested query segmentation. Also, the experiments for queries (**Q500+Q700**) were conducted separately from **S300** and **QRand**. Thus, we had six HITs in all. The concept of flat and nested segmentation was introduced to the Turkers with the help of examples presented in two short videos[6]. When in doubt regarding the meaning of a query, the Turkers were advised to issue the query on a search engine of their choice and find out its possible interpretation(s). Note that we intentionally kept definitions of flat and nested segmentation fuzzy because (a) it would require very long instruction manuals to cover all possible cases and (b) Turkers do not tend to read verbose and complex instructions. Table 3 summarizes other specifics of HITs.

Honey pots or trap questions whose answers are known a priori are often included in a HIT to identify turkers who are unable to solve the task appropriately leading to incorrect annotations. However, this trick cannot be employed in our case because there is no notion of an absolutely correct segmentation. We observe that even with unambiguous queries, even expert annotators may disagree on some of the segment boundaries. Hence, we decided to include annotations from all the

[6] Flat: `http://youtu.be/eMeLjJIvIh0`, Nested: `http://youtu.be/xE3rwANbFvU`

turkers, except for those that were syntactically ill-formed (e.g., non-binary nested segmentation).

4 Inter Annotator Agreement

Inter Annotator Agreement is the only way to judge the reliability of annotated data in absence of an end application. Therefore, before we can venture into analysis of the experimental data, we need to formalize the notion of IAA for flat and nested queries. The task is non-trivial for two reasons. First, traditional IAA measures are defined for a fixed set of annotators. However, for crowdsourcing based annotations, different annotators might have annotated different parts of the dataset. For instance, we observed that a total of 128 turkers have provided the flat annotations for **Q700**, when we had only asked for 10 annotations per query. Thus, on average, a turker has annotated only 7.81% of the 700 queries. In fact, we found that 31 turkers had annotated less than 5 queries. Hence, measures such as Cohen's κ (1960) cannot be directly applied in this context because for crowdsourced annotations, we cannot meaningfully compute annotator-specific distribution of the labels and biases.

Second, most of the standard annotation metrics do not generalize for flat segmentation and trees. Artstein and Poesio (2008) provides a comprehensive survey of the IAA metrics and their usage in NLP. They note that all the metrics assume that a fixed set of labels are used for items. Therefore, it is far from obvious how to compare chunking or segmentation that *covers* the whole text or that might have *overlapping* units as in the case of nested segmentation. Furthermore, we would like to compare the reliability of flat and nested segmentation, and therefore, ideally we would like to have an IAA metric that can be meaningfully applied to both of these cases.

After considering various measures, we decided to appropriately generalize one of the most versatile and effective IAA metrics proposed till date, the Kripendorff's α (2004). To be consistent with prior work, we will stick to the notation used in Artstein and Poesio (2008) and redefine the α in the context of flat and nested segmentation. Note that though the notations introduced here will be from the perspective of queries, it is equally applicable to sentences and the generalization is straightforward.

4.1 Notations and Definitions

Let Q be the set of all queries with cardinality $\mathbf{q}$. A query $q \in Q$ can be represented as a sequence of $|q|$ words: $w_1 w_2 \ldots w_{|q|}$. We introduce $|q-1|$ random variables, $b_1, b_2, \ldots b_{|q|-1}$, such that b_i represents the boundary between the words w_i and w_{i+1}. A flat or nested segmentation of q, represented by q_j, j varying from 1 to total number of annotations $\mathbf{c}$, is a particular instantiation of these boundary variables as described below.

Definition. A *flat segmentation*, q_j can be uniquely defined by a binary assignment of the boundary variables $b_{j,i}$, where $b_{j,i} = 1$ iff w_i and w_{i+1} belong to two different flat segments. Otherwise, $b_{j,i} = 0$. Thus, q has $2^{|q|-1}$ possible flat segmentations.

Definition. A *nested segmentation* q_j can also be uniquely defined by assigning non-negative integers to the boundary variables such that $b_{j,i} = 0$ iff words w_i and w_{i+1} form an atomic segment (i.e., they are grouped together), else $b_{j,i} = 1 + max(left_i, right_i)$, where $left_i$ and $right_i$ are the heights of the largest subtrees ending at w_i and beginning at w_{i+1} respectively.

This numbering scheme for nested segmentation can be understood through Fig. 1. Every internal node of the binary tree corresponding to the nested segmentation is numbered according to its height. The lowest internal nodes, both of whose children are query words, are assigned a value of 0. Other internal nodes get a value of one greater than the height of its higher child. Since every internal node corresponds to a boundary, we assign the height of the node to the corresponding boundaries. The number of unique nested segmentations of a query of length $|q|$ is the $(|q|-1)^{th}$ *Catalan number*[7].

Boundary variables for flat and nested segmentation are illustrated with an example of each kind in Tables 1 and 2 (last column).

4.2 Krippendorff 's α for Segmentation

Krippendorff 's α (Krippendorff, 2004) is an extremely versatile agreement coefficient, which is based on the assumption that the expected agreement is calculated by looking at the overall distribution of judgments without regard to which annotator produced them (Artstein and Poesio, 2008). Hence, it is appropriate for crowdsourced annotation, where the judgments come from a large num-

[7] http://goo.gl/vKQvK

ber of unrelated annotators. Moreover, it allows for different magnitudes of disagreement, which is a useful feature as we might want to differentially penalize disagreements at various levels of the tree for nested segmentation.

α is defined as

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{s^2_{within}}{s^2_{total}} \quad (1)$$

where D_o and D_e are, respectively, the observed and expected disagreements that are measured by s^2_{within} – the variance within the annotation of an item and s^2_{total} – variance across annotations of all items. We adapt the equations presented in pp.565-566 of Artstein and Poesio (2008) for measuring these quantities for queries:

$$s^2_{within} = \frac{1}{2\mathbf{qc}(\mathbf{c}-1)} \sum_{q \in Q} \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} d(q_m, q_n) \quad (2)$$

$$s^2_{total} = \frac{1}{2\mathbf{qc}(\mathbf{qc}-1)} \sum_{q \in Q} \sum_{m=1}^{\mathbf{c}} \sum_{q' \in Q} \sum_{n=1}^{\mathbf{c}} d(q_m, q'_n) \quad (3)$$

where, $d(q_m, q'_n)$ is a distance metric for the agreement between annotations q_m and q'_n.

We define two different distance metrics d_1 and d_2 that are applicable to flat and nested segmentation. We shall first define these metrics for comparing queries with equal length (i.e., $|q| = |q'|$):

$$d_1(q_m, q'_n) = \frac{1}{|q|-1} \sum_{i=1}^{|q|-1} |b_{m,i} - b'_{n,i}| \quad (4)$$

$$d_2(q_m, q'_n) = \frac{1}{|q|-1} \sum_{i=1}^{|q|-1} |b^2_{m,i} - (b'_{n,i})^2| \quad (5)$$

While d_1 penalizes all disagreements equally, d_2 penalizes disagreements higher up the tree more. d_2 might be a desirable metric for nested segmentation, because research on sentence chunking shows that annotators agree more on clause or major phrase boundaries, even though they may not always agree on intra-clausal or intra-phrasal boundaries (Bali et al., 2009). Note that for flat segmentation, d_1 and d_2 are identical, and hence we will denote them as d.

We propose the following extension to these metrics for queries of unequal lengths. Without loss of generality, let us assume that $|q| < |q'|$. k is 1 or 2; $r = |q'| - |q| + 1$.

$$d_k(q_m, q'_n) = \frac{1}{r(|q|-1)} \sum_{a=0}^{r-1} \sum_{i=1}^{|q|-1} |b^k_{m,i} - (b'_{n,i+a})^k| \quad (6)$$

4.3 IAA under Random Bias Assumption

Krippendorff's α uses the cross-item variance as an estimate of chance agreement, which is reliable in general. However, this might result in misleadingly low values of IAA, especially when the items in the set are indeed expected to have similar annotations. To resolve this, we also compute the chance agreement under a random bias model. The random model assumes that *all the structural annotations of q are equiprobable*. For flat segmentation, it boils down to the fact that all the $2^{|q|-1}$ annotations are equally likely, which is equivalent to the assumption that any boundary variable b_i has 0.5 probability of being 0 and 0.5 for 1.

Analytical computation of the expected probability distributions of $d_1(q_m, q_n)$ and $d_2(q_m, q_n)$ is harder for nested segmentation. Therefore, we programmatically generate all possible trees for q, which is again dependent only on $|q|$ and compute d_1 and d_2 between all pairs of trees, from which the expected distributions can be readily estimated. Let us denote this expected cumulative probability distribution for flat segmentation as $P_d(x; |q|)$ = the probability that for a pair of randomly chosen flat segmentations of q, q_m and q_n, $d(q_m, q_n) \geq x$. Likewise, let $P_{d_1}(x; |q|)$ and $P_{d_2}(x; |q|)$ be the respective probabilities that for any two nested segmentations q_m and q_n of q, the following holds: $d_1(q_m, q_n) \geq x$ and $d_2(q_m, q_n) \geq x$.

We define the IAA under random bias model as (k is 1, 2 or null):

$$S = \frac{1}{\mathbf{qc}^2} \sum_{q \in Q} \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} P_{d_k}(d_k(q_m, q_n); |q|) \quad (7)$$

Thus, S is the expected probability of observing a similar or worse agreement by random chance, averaged over all pairs of annotations for all queries, and not a chance corrected IAA metric such as α. Thus, $S = 1$ implies that the observed agreement is *almost always better than* that by random chance and $S = 0.5$ and 0 respectively imply that the observed agreement is *as good as* and *almost always worse than* that by random chance. We also note that a high value of S and low value

Dataset	Flat	Nested	
	d_1	d_1	d_2
Q700	0.21(0.59)	0.21(0.89)	0.16(0.68)
Q500	0.22(0.62)	0.15(0.70)	0.15(0.44)
QG500	0.61(0.88)	0.66(0.88)	0.67(0.80)
S300	0.27(0.74)	0.18(0.94)	0.14(0.75)
U250	0.23(0.89)	0.42(0.90)	0.30(0.78)
B250	0.22(0.86)	0.34(0.88)	0.22(0.71)
T250	0.20(0.86)	0.44(0.89)	0.34(0.76)

Table 4: Agreement Statistics: $\alpha(S)$.

of α indicate that though the annotators agree on the judgment of individual items, they also tend to agree on judgments of two different items, which in turn, could be due to strong annotator biases or due to lack of variability of the dataset.

In the supplementary material, computations of α and S have been explained in further details through worked out examples. Tables for the expected distributions of d, d_1 and d_2 under the random annotation assumption are also available.

5 Results

Table 4 reports the values of α and S for flat and nested segmentation on the various datasets. For nested segmentation, the values were computed for two different distance metrics d_1 and d_2. As expected, the highest value of α for both flat and nested segmentation is observed for gold annotations. An $\alpha > 0.6$ indicates quite good IAA, and thus, reliable annotations. Higher α for nested segmentation **QG500** than flat further validates our initial postulate that nested segmentation may reduce disagreement from granularity issues inherent in the definition of flat segmentation.

Opposite trends are observed for **Q700**, **Q500** and **S300**, where α for flat is the highest, followed by that for nested using d_1, and then d_2. Moreover, except for flat segmentation of sentences, α lies between 0.14 and 0.22, which is quite low. This clearly shows that segmentation, either flat or nested, cannot be reliably procured through crowdsourcing. Lower α for d_2 than d_1 further indicates that annotators disagree more for higher levels of the trees, contrary to what we had expected. However, nearly equal IAA for sentences and queries implies that low agreement may not be an outcome of inherent ambiguity in the structure of queries. Slightly higher α for flat segmentation and a much higher α for nested segmentation of **QRand** reinforce the fact that low IAA is not due to a lack of structure in queries.

It is interesting to note that α for nested segmentation of **S300** and all segmentations of **QRand** are low or medium despite the fact that S is very high in all these cases. Thus, it is clear that annotators have a strong bias towards certain structures across queries. In the next section, we will analyze some of these biases. We also computed the IAA between **QG500** and **Q500**, and found $\alpha = 0.27$. This is much lower than α for **QG500**, though slightly higher than that for **Q500**. We did not observe any significant variation in agreement with respect to the length of the queries.

6 Biases in Annotation

The IAA statistics clearly show that there are certain strong biases in both flat and nested query segmentation, especially those obtained through crowdsourcing. To identify these biases, we went through the annotations and came up with possible hypotheses, which we tried to verify through statistical analysis of the data. Here, we report the most prominent biases that were thus discovered.

Bias 1: *During flat segmentation, annotators prefer dividing the query into two segments of roughly equal length.*

As discussed earlier, one of the major problems of flat segmentation is the fuzziness in granularity. In our experiments, we intentionally left the decision of whether to go for fine or coarse-grained segmentation to the annotator. However, it is surprising to observe that annotators typically divide the query into two segments (see Fig. 3, plots A1 and A2), and at times three, but hardly ever more than three. This bias is observed across queries, sentences and random queries, where the percentage of annotations with 2 or 3 segments are greater than 83%, 91% and 96% respectively. This bias is most strongly visible for **QRand** because the lack of syntactic or semantic cohesion between the words provides no clue for segmentation.

Furthermore, we observe that typically segments tend to be of equal length. For this, we computed standard deviations (sd) of segment lengths for all annotations having 2 or 3 segments; the distribution of sd is shown in Fig. 3, plots B1 and B2. We observe that for all datasets, sd lies mainly between 0.5 and 1 (for perspective, consider a query with 7 words; with two segments of length 3 and

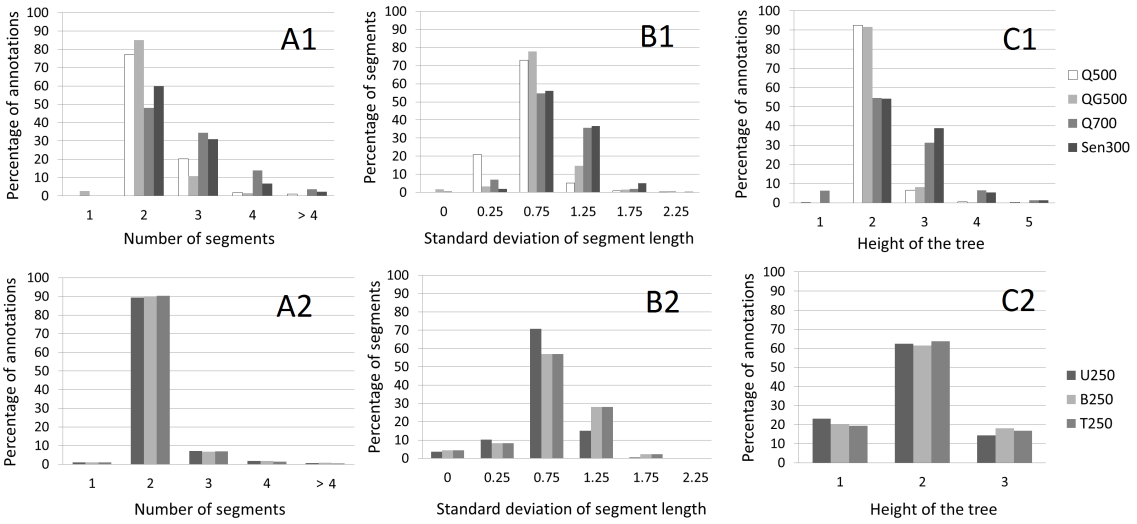

Figure 3: Analysis of annotation biases: A1, A2 – number of segments per flat segmentation vs. length; B1, B2 – standard deviation of segment length for flat segmentation; C1, C2 – distribution of the tree heights in nested segmentation.

Length	Expected	Q500	QG500	Q700	S300	QRand
5	2.57	2.00	2.02	2.08	2.02	2.01
6	3.24	2.26	2.23	2.23	2.24	2.02
7	3.88	2.70	2.71	2.67	2.55	2.62
8	4.47	2.89	2.68	2.72	2.72	2.35

Table 5: Average height for nested segmentation.

4 the sd is 0.5, and for 2 and 5, the sd is 1.5), implying that segments are roughly of equal length.

It is likely that due to this bias, the S or observed agreement is moderately high for queries and very high for sentences, but then it also leads to high agreement across different queries and sentences (i.e., high s^2_{total}) especially when they are of equal length, which in turn brings down the value of α – the true agreement after bias correction.

Bias 2: *During nested segmentation, annotators prefer balanced binary trees.*

Quite analogous to bias 1, for nested segmentation we observe that annotators tend to prefer more balanced binary trees. Fig. 3 plots C1 and C2 show the distribution of the tree heights for various cases and Table 5 reports the corresponding average height of the trees for queries and sentences of various lengths and the the expected value of the height if all trees were equally likely. The observed heights are much lower than the expected values clearly implying the preference of the annotators for more balanced trees.

Thus, the crowd seems to choose the middle path, avoiding extremes and hence may not be a reliable source of annotation for query segmentation. It can be argued that similar biases are also observed for gold annotations, and therefore, probably it is the inherent structure of the queries and sentences that lead to such biased distribution of segmentation patterns. However, note that α for **QG500** is much higher than all other cases, which shows that the true agreement between gold annotators is immune to such biases or skewed distributions in the datasets. Furthermore, high values of α for **QRand** despite the very strong biases in annotation shows that there perhaps is very little choice that the annotators have while segmenting randomly generated queries. On the other hand, the textual coherence of the real queries and sentences provide many different choices for segmentation and the Turker typically gets carried away by these biases, leading to low α.

Bias 3: *Phrase structure drives segmentation only when reconcilable with Bias 1.* Whenever the sentence or query has a verb phrase (VP) spanning roughly half of it, annotators seem to chunk before the VP as one would expect, quite as often as just after the verb, which is quite unexpected. For instance, the sentence `A gentle sarcasm ruffled her anger.` gathers as many as eight flat annotations with a boundary between `sarcasm` and `ruffled`, and four with a boundary between `ruffled` and `her`. However, if the VP is very short consisting of a single verb, as in `A fleeting and furtive air`

Position	Q500	QG500	Q700	S300	QRand
Both	2.24	0.37	2.78	2.08	0.63
None	50.34	56.85	35.74	35.84	39.81
Right	23.86	21.50	19.02	12.52	15.23
Left	18.08	15.97	40.59	45.96	21.21

Table 6: Percentages of positions of segment boundaries with respect to prepositions. Prepositions occurring in the beginning or end of a query/sentence have been excluded from the analysis; hence, numbers in a column do not total 100.

`of triumph erupted.`, annotators seem to attempt for a balanced annotation due to **Bias 1**. As a clear middle boundary is not present in such sentences, the annotations show a lot more variation and disagreement. For instance, only 1 out of 10 annotations had a boundary before `erupted` in the above example. In fact, at least one annotation had a boundary after each word in the sentence, with no clear majority.

Bias 4: *Prepositions influence segment boundaries differently for queries and sentences.* We automatically labeled all the prepositions in the flat annotations and classified them according to the criterion of whether a boundary was placed immediately before or after it, or on both sides or neither side. The statistics, reported in Table 6, show that for NL sentences a majority of the boundaries are present before the preposition, marking the beginning of a prepositional phrase. However, for queries, a much richer pattern emerges depending on the specific preposition. For instance, `to`, `of` and `for` are often chunked with the previous word (e.g., `how to | choose a bike size`, `birthday party ideas for | one year old`). We believe that this difference is because in sentences due to the presence of a verb, the PP has a well-defined head, lack of which leads to preposition in queries getting chunked with words that form more commonly seen patterns (e.g., `flights to` and `tickets for`).

Bias 3 and 4 present the complex interpretation of the structure of queries by the annotators which could be due to some emerging cognitive model of queries among the search engine users. This is a fascinating and unexplored aspect of query structures that demands deeper investigation through cognitive and psycholinguistic experiments.

7 Conclusion

We have studied various aspects of query segmentation through crowdsourcing by designing and conducting suitable experiments. Analysis of experimental data leads us to conclude the following: (a) crowdsoucing may not be a very effective way to collect judgments for query segmentation; (b) addressing fuzziness of granularity for flat segmentation by introducing strict binary nested segments does not lead to better agreement in crowdsourced annotations, though it definitely improves the IAA for gold standard segmentations, implying that low IAA in flat segmentation among experts is primarily an effect of unspecified granularity of segments; (c) low IAA is not due to the inherent structural ambiguity in queries as this holds true for sentences as well; (d) there are strong biases in crowdsourced annotations, mostly because turkers prefer more balanced segment structures; and (e) while annotators are by and large guided by linguistic principles, application of these principles differ between query and NL sentences and also closely interact with other biases.

One of the important contributions of this work is the formulation of a new IAA metric for comparing across flat and nested segmentations, especially for crowdsourcing based annotations. Since trees are commonly used across various linguistic annotations, this metric can have wide applicability. The metric, moreover, can be easily adapted to other annotation schemes as well by defining an appropriate distance metric between annotations. Since large scale data for query segmentation is very useful, it would be interesting to see if the problem can be rephrased to the Turkers in a way so as to obtain more reliable judgments. Yet a deeper question is regarding the theoretical status of query structure, which though in an emergent state is definitely an operating model for the annotators. Our future work in this area would specifically target understanding and formalization of the theoretical model underpinning a query.

Acknowledgments

We thank Ed Cutrell and Andrew Cross, Microsoft Research Lab India, for their help in setting up the AMT experiments. We would also like to thank Anusha Suresh, IIT Kharagpur, India, for helping us with data preparation.

References

Steven P. Abney. 1991. *Parsing By Chunks*. Kluwer Academic Publishers.

Steven P. Abney. 1992. Prosodic Structure, Performance Structure And Phrase Structure. In *Proceedings 5th DARPA Workshop on Speech and Natural Language*, pages 425–428. Morgan Kaufmann.

Steven P. Abney. 1995. Chunks and dependencies: Bringing processing evidence to bear on syntax. *Computational Linguistics and the Foundations of Linguistic Theory*, pages 145–164.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Kalika Bali, Monojit Choudhury, Diptesh Chatterjee, Sankalan Prasad, and Arpit Maheswari. 2009. Correlates between Performance, Prosodic and Phrase Structures in Bangla and Hindi: Insights from a Psycholinguistic Experiment. In *Proceedings of International Conference on Natural Language Processing*, pages 101 – 110.

Michael Bendersky, W. B. Croft, and David A. Smith. 2009. Two-stage query segmentation for information retrieval. In *Proceedings of the 32nd international ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pages 810–811. ACM.

Shane Bergsma and Qin Iris Wang. 2007. Learning Noun Phrase Query Segmentation. In *Proceedings of Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 819–826.

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, and KV Ramakrishnamacharyulu. 1995. *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi.

David J. Brenes, Daniel Gayo-Avello, and Rodrigo Garcia. 2010. On the fly query segmentation using snippets. In *CERI '10*, pages 259–266.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 286–295. Association for Computational Linguistics.

Vitor R Carvalho, Matthew Lease, and Emine Yilmaz. 2011. Crowdsourcing for search evaluation. *ACM Sigir forum*, 44(2):17–22.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Zhicheng Dou, Ruihua Song, Xiaojie Yuan, and Ji-Rong Wen. 2008. Are Click-through Data Adequate for Learning Web Search Rankings? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 73–82. ACM.

Matthias Hagen, Martin Potthast, Benno Stein, and Christof Bräutigam. 2011. Query Segmentation Revisited. In *Proceedings of the 20th International Conference on World Wide Web*, pages 97–106. ACM.

Matthias Hagen, Martin Potthast, Anna Beyer, and Benno Stein. 2012. Towards Optimum Query Segmentation: In Doubt Without. In *Proceedings of the Conference on Information and Knowledge Management*, pages 1015–1024.

Jian Huang, Jianfeng Gao, Jiangbo Miao, Xiaolong Li, Kuansan Wang, Fritz Behr, and C. Lee Giles. 2010. Exploring web scale language models for search query processing. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 451–460, New York, NY, USA. ACM.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to its Methodology*. Sage,Thousand Oaks, CA.

Yanen Li, Bo-Jun Paul Hsu, ChengXiang Zhai, and Kuansan Wang. 2011. Unsupervised query segmentation using clickthrough for information retrieval. In *SIGIR '11*, pages 285–294. ACM.

Knut Magne Risvik, Tomasz Mikolajewski, and Peter Boros. 2003. Query segmentation for web search. In *WWW (Posters)*.

Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury, and Srivatsan Laxman. 2012. An IR-based Evaluation Framework for Web Search Query Segmentation. In *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pages 881–890. ACM.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bin Tan and Fuchun Peng. 2008. Unsupervised Query Segmentation Using Generative Language Models and Wikipedia. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, pages 347–356. ACM.

Chao Zhang, Nan Sun, Xia Hu, Tingzhu Huang, and Tat-Seng Chua. 2009. Query segmentation based on eigenspace similarity. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 185–188, Stroudsburg, PA, USA. Association for Computational Linguistics.