# Identifying Well-formed Natural Language Questions

**Manaal Faruqui** **Dipanjan Das**
Google AI Language

## Abstract

Understanding search queries is a hard problem as it involves dealing with "word salad" text ubiquitously issued by users. However, if a query resembles a well-formed question, a natural language processing pipeline is able to perform more accurate interpretation, thus reducing downstream compounding errors. Hence, identifying whether or not a query is well formed can enhance query understanding. Here, we introduce a new task of identifying a well-formed natural language question. We construct and release a dataset of 25,100 publicly available questions classified into well-formed and non-wellformed categories and report an accuracy of 70.7% on the test set. We also show that our classifier can be used to improve the performance of neural sequence-to-sequence models for generating questions for reading comprehension.

## 1 Introduction

User issued search queries often do not follow formal grammatical structure, and require specialized language processing (Bergsma and Wang, 2007; Barr et al., 2008; Manshadi and Li, 2009; Mishra et al., 2011). Traditional natural language processing (NLP) tools trained on formal text (e.g. treebanks) often have difficulty analyzing search queries; the lack of regularity in the structure of queries makes it difficult to train models that can optimally process the query to extract information that can help understand the user intent behind the query (Baeza-Yates et al., 2006).

One clear direction to improve query processing is to annotate a large number of queries with the desired annotation scheme. However, such an annotation can be prohibitively expensive and models trained on such queries might suffer from freshness issues, as the domain and nature of queries evolve frequently (Markatos, 2001; Bawa et al., 2003; Roy et al., 2012). Another direction is to obtain a paraphrase of the given query that is a grammatical natural language question, and then analyze that paraphrase to extract the required information (Nogueira and Cho, 2017; Buck et al., 2018). There are available tools and datasets, such as Quora question paraphrases and the Paralex dataset (Fader et al., 2013) – for identifying query paraphrases (Wang et al., 2017; Tomar et al., 2017), but these datasets do not contain information about whether a query is a natural language question or not. Identifying well-formed natural language questions can also facilitate a more natural interaction between a user and a machine in personal assistants or chatbots (Yang et al., 2014; Mostafazadeh et al., 2016) or while recommending related queries in search-engines.

Identifying a well-formed question should be easy by parsing with a grammar, such as the English resource grammar (Copestake and Flickinger, 2000), but such grammars are highly precise and fail to parse more than half of web queries. Thus, in this paper we present a model to predict whether a given query is a well-formed natural language question. We construct and publicly release a dataset of 25,100 queries annotated with the probability of being a well-formed natural language question (§2.1). We then train a feed-forward neural network classifier that uses the lexical and syntactic features extracted from the query on this data (§2.2). On a test set of 3,850 queries, we report an accuracy of 70.1% on the binary classification task. We also demonstrate that such a query well-formedness classifier can be used to improve the quality of a sequence-to-sequence question generation model (Du et al., 2017) by showing an improvement of 0.2 BLEU score in its performance (§3). Our dataset ise available for download at `http://goo.gl/language/query-wellformedness`.



*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 798–803
Brussels, Belgium, October 31 - November 4, 2018. 

| Query | Well-formed? | Reasoning |
|---|---|---|
| what is the breed of scooby doo? | 1 | Grammatical and an explicit question |
| tell me whats the breed of scooby doo? | 0 | A command but not a question |
| headache evenings? | 0 | Ungrammatical and not a question |
| what causes headaches during evenings | 1 | Grammatical and an explicit question |
| what 12.5 as a fraction? | 0 | An explicit question but ungrammatical |

Table 1: Examples of well-formed and non-wellformed queries according to the annotation guideline.

## 2 Well-formed Natural Language Question Classifier

In this section we describe the data annotation, and the models used for question well-formedness classification.

### 2.1 Dataset Construction

We use the Paralex corpus (Fader et al., 2013) that contains pairs of noisy paraphrase questions. These questions were issued by users in WikiAnswers (a Question-Answer forum) and consist of both web-search query like constructs (*"5 parts of chloroplast?"*) and well-formed questions (*"What is the punishment for grand theft?"*), and thus is a good resource for constructing the question well-formedness dataset. We select 25,100 queries from the unique list of queries extracted from the corpus such that no two queries in the selected set are paraphrases. The queries are then annotated into well-formed or non-wellformed questions. We define a query to be a well-formed natural language question if it satisfies the following:

1. Query is grammatical.
2. Query is an explicit question.
3. Query does not contain spelling errors.

Table 1 shows some examples that were shown to the annotators to illustrate each of the above conditions. Every query was labeled by five different crowdworkers with a binary label indicating whether a query is well-formed or not. We average the ratings of the five annotators to get the probability of a query being well-formed. Table 2.1 shows some queries with obtained human annotation. Humans are pretty good at identifying an implicit query (*"Population of owls..."*) or a simple well-formed question (*"What is released..."*), but may miss out on subtle spelling mistakes like "disscovered" or disagree on whether the determiner "the" is needed before the word "genocide" (*"What countries have genocide happened in?"*). Similar to other NLP tasks like entailment (Dagan et al., 2006; Bowman et al., 2015), paraphrasing (Wieting et al., 2015) etc. we rely on the wisdom of the crowd to get such annotations in order to make the data collection scalable and language-independent.

| Query ($q$) | $p_{wf}(q)$ |
|---|---|
| population of owls just in north america? | 0.0 |
| who disscoverd rihanna? | 0.2 |
| what countries have genocide happened in? | 0.6 |
| what is released when an ion is formed? | 1.0 |

Table 2: Examples of human annotations on query well-formedness.

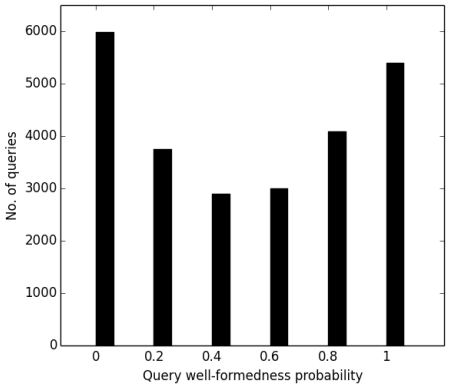


Figure 1: The distribution of the annotated questions according to well-formedness probability.

Figure 1 is the histogram of query well-formedness probability across the dataset. Interestingly, the number of queries where at least 4 or more annotators agree[1] on well-formedness is large: $|\{q \mid 0.8 \leq p_{wf}(q) \leq 0.2\}| = 19206$ queries. These constitute 76.5% of all queries in the dataset. The Fleiss' kappa (Fleiss, 1971) for measuring agreement among multiple annotators is computed to be $\kappa = 0.52$ which shows moderate agreement (Landis and Koch, 1977). We

[1] At least 4 annotators label the query with 0 or 1.

then randomly divided the dataset in approx. 70%, 15%, 15% ratio into training, development and test sets containing 17500, 3750, and 3850 queries respectively. While testing, we consider every query well-formed where at least 4 out of 5 annotators ($p_{wf} \geq 0.8$) marked it as well-formed.[2]

### 2.2 Model

We use a feed-forward neural network with 2 hidden layers with ReLU activations (Glorot et al., 2011) on each layer and a softmax at the output layer predicting 0 or 1. We extract a variety of features from the query which can be helpful in the classification. We extract character-3, 4-grams and word-1, 2-grams as they can be helpful in capturing spelling errors. In addition to lexical features, we also extract syntactic features that can inform the model on any anomaly in the structure of the query. Specifically, we annotate the query with POS-tags using SyntaxNet POS tagger (Alberti et al., 2015) and extract POS-1, 2, 3-grams.[3] Every feature in the network is represented as a real-valued embedding. All the $n$-grams embeddings of every feature type are summed together and concatenated to form the input layer as shown in Figure 2. The model is trained using cross-entropy loss against the gold labels for each query. The hyperparameters are tuned to maximize accuracy on the dev set and results are reported on the test set.

**Hyperparameters.** We fix the size of the first and second hidden layers to be 128 and 64 respectively. The character n-gram embeddings were of length 16 and all other feature embeddings were of length 25. We use stochastic gradient descent with momentum for optimization with learning rate tuned over $[0.001 - 0.3]$, a batch size of 32 and 50000 training steps.

### 2.3 Experiments

**Baselines.** The majority class baseline is 61.5% which corresponds to all queries being classified non-wellformed. The question word baseline that classifies any query starting with a question word as a well-formed question gets 54.9%.[4] Also, we used a single-layer word-level biLSTM encoder with hidden layer of length 50 to encode the question and then use this representation in the softmax layer to predict the label (Lee and Dernoncourt, 2016). This classifier achieved 65.8%.

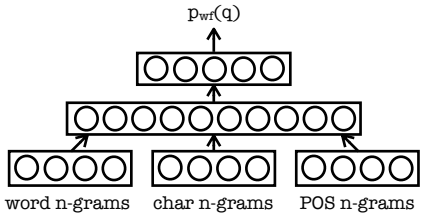


Figure 2: A feed-forward neural network for query well-formedness classification.

| Model | Accuracy (%) |
|---|---|
| majority class baseline | 61.5 |
| word bi-LSTM baseline | 65.8 |
| question word baseline | 54.9 |
| word-1 | 65.4 |
| word-1, 2 | 65.5 |
| word-1, 2 char-3, 4 | 66.9 |
| word-1, 2 POS-1, 2, 3 | **70.7** |
| word-1, 2 char-3, 4 POS-1, 2, 3 | 70.2 |
| Approx. human upper bound | 88.4 |

Table 3: Performance of well-formedness query classifier on the test set.

**Results.** The best performance obtained is 70.7% while using word-1, 2-grams and POS-1, 2, 3-grams as features. Using POS $n$-grams gave a strong boost of 5.2 points over word unigrams and bigrams. Although character-3, 4-grams gave improvement over word unigrams and bigrams, the performance did not sustain when combined with POS tags.[5] A random sample of 1000 queries from the test set were annotated by one of the authors of the paper with proficiency in English, which matched the gold label with 88.4% accuracy providing an approximate upper-bound for model performance.

A major source of error is our model's inability to understand deep semantics and syntax. For example, *"What is the history of dirk bikes?"* is labeled as a non-wellformed question with $p_{wf} =$

[2] We randomly selected 100 queries and manually determined if each of those queries were well-formed. We found $p_{wf}(q) = 0.8$ to be the value above which all queries were well-formed.

[3] The use of dependency labels as features and use of pre-trained Glove embeddings did not show improvement and hence omitted here for space constraints.

[4] List of question words: https://en.wikipedia.org/wiki/Interrogative_word

[5] We assumed character $n$-grams to help identify spelling mistakes, but our dataset has relatively few misspelled words–only 6 in 100 random queries.

0 by annotators because of the misspelled word "dirk" (the correct word is "dirt"). However, the POS tagger identifies "dirk" as a noun and as "NN NNS" is a frequent POS-bigram, our model tags it as a well-formed question with $p_{wf} = 0.8$, unable to identify that the word does not fit in the context of the question. Another source of error is the inability to capture long term grammatical dependencies. For example, in *"What sort of work did Edvard Munch made ?"* the verb "made" is incorrectly in the past tense instead of present tense. Our model is unable to capture the relationship between "did" and "made" and thus marks this as a well-formed question.

## 3 Improving Question Generation

Automatic question generation is the task of generating questions that ask about the information or facts present in either a given sentence or paragraph (Vanderwende, 2008; Heilman and Smith, 2010). Du et al. (2017) present a state-of-the-art neural sequence-to-sequence model to generate questions from a given sentence/paragraph. The model used is an attention-based encoder-decoder network (Bahdanau et al., 2015), where the encoder reads in a given text and the decoder is an LSTM RNN that produces the question by predicting one word at a time.

Du et al. (2017) use the SQuAD question-answering dataset (Rajpurkar et al., 2016) to develop a question generation dataset by pairing sentences from the text with the corresponding questions. The question generation dataset contains approx 70k, 10k, and 12k training, development and test examples. Their current best model selects the top ranked question from the $n$-best list produced by the decoder as the output. We augment their system by training a discriminative reranker (Collins and Koo, 2005) with the model score of the question generation model and the well-formedness probability of our classifier as features to optimize BLEU score (Papineni et al., 2002) between the selected question from the 10-best list and the reference question on the development set. We then use this reranker to select the best question from the 10-best list of the test set.

We use the evaluation package released by Chen et al. (2015) to compute BLEU-1 and BLEU-4 scores.[6] Table 4 shows that the reranked question selected using our query well-formedness classifier improves the BLEU-4 score of a seq-to-seq question generation model from 12.0 to 12.2. The oracle improvement, by selecting the sentence from the list that maximizes the BLEU-4 score is 15.2. However, its worth noting that an increase in well-formedness doesn't guarantee an improved BLEU score, as the oracle sentence maximizing the BLEU score might be fairly non-wellformed (Callison-Burch et al., 2006). For example, *"who was elected the president of notre dame in?"* has a higher BLEU score to the reference *"who was the president of notre dame in 1934?"* than our well-formed question *"who was elected the president of notre dame?"*. Figure 3 shows a question generation example with the output of Du et al. (2017) as the baseline result and the reranked question using the wellformed probability.

[6] BLEU-x uses precision computed over $[1, x]$-grams.

| Model | BLEU-1 | BLEU-4 |
|---|---|---|
| Baseline | 41.3 | 12.0 |
| Reranked | **41.6** | **12.2** |

Table 4: Reranking the $n$-best output of a neural seq2seq question generation model using well-formedness probability.

| |
|---|
| **Sentence:** montana is home to the rocky mountain elk foundation and has a historic big game hunting tradition.<br>**Gold question:** what is the name of the big game hunting foundation in montana? |
| **seq2seq:** what is a historic big game hunting tradition? ($p_{wf} = 0.7$)<br>**Reranked:** what is the name of the historic big game tradition? ($p_{wf} = 0.8$) |

Figure 3: Example showing question selection from the $n$-best list using our reranking model.

## 4 Related Work

We have referenced much of the related work throughout the paper. We now review another orthogonally related field of work. Grammatical error correction (GEC) is the task of correcting the grammatical errors (if any) in a piece of text (Ng et al., 2014). As GEC includes not just identification of ungrammatical text but also correcting the text to produce grammatical text, its a more complex task. However, grammatical error prediction (Schmaltz et al., 2016; Daudaravicius et al., 2016) is the task of classifying whether or not a sentence is grammatical, which is more closely related to

our task as classifying a question as well-formed requires making judgement on both the style and grammar of the text.

## 5 Conclusion

We proposed a new task of well-formed natural language question identification and established a strong baseline on a new dataset that can be downloaded at: http://goo.gl/language/query-wellformedness. We also showed that question well-formedness information can be a helpful signal in improving state-of-the-art question generation systems.

## References


Chris Alberti, David Weiss, Greg Coppola, and Slav Petrov. 2015. Improved transition-based parsing and tagging with neural networks. In *Proc. of EMNLP*.

Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. 2006. The intention behind web queries. In *Proc. of SPIRE*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Cory Barr, Rosie Jones, and Moira Regelson. 2008. The linguistic structure of English web-search queries. In *Proc. of EMNLP*.

Mayank Bawa, Roberto J. Bayardo, Jr., Sridhar Rajagopalan, and Eugene J. Shekita. 2003. Make it fresh, make it quick: Searching a network of personal webservers. In *Proc. of WWW*.

Shane Bergsma and Qin Iris Wang. 2007. Learning noun phrase query segmentation. In *Proc. of EMNLP-CoNLL*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Andrea Gesmundo, Neil Houlsby, Wojciech Gajewski, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *Proc. of ICLR*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proc. of EACL*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Comput. Linguist.*, 31(1):25–70.

Ann A Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage english grammar using HPSG. In *Proc. of LREC*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges.*, pages 177–190. Springer.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A report on the automatic evaluation of scientific writing shared task. In *Proc. of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proc. of ACL*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proc. of ACL*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proc. of AISTATS*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Proc. of NAACL*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proc. of NAACL*.

Mehdi Manshadi and Xiao Li. 2009. Semantic tagging of web search queries. In *Proc. of ACL-IJCNLP*.

Evangelos P. Markatos. 2001. On caching search engine query results. *Computer Communications*, 24(2):137–143.

Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly, Srivatsan Laxman, and Monojit Choudhury. 2011. Unsupervised query segmentation using only query logs. In *Proc. of WWW*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proc. of ACL*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *CoNLL Shared Task*.

Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. In *Proc. of EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. of ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.

Rishiraj Saha Roy, Monojit Choudhury, and Kalika Bali. 2012. Are web search queries an evolving protolanguage. In *In Proc. of Evolang*.

Allen Schmaltz, Yoon Kim, Alexander M. Rush, and Stuart Shieber. 2016. Sentence-level grammatical error identification as sequence-to-sequence correction. In *Proc. of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*.

Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural paraphrase identification of questions with noisy pretraining. In *Proc. of First Workshop on Subword and Character Level Models in NLP*.

Lucy Vanderwende. 2008. The importance of being important: Question generation. In *Proc. of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proc. of IJCAI*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *TACL*, 3:345–358.

Jie Yang, Claudia Hauff, Alessandro Bozzon, and Geert-Jan Houben. 2014. Asking the right question in collaborative QA systems. In *Proc. of Conference on Hypertext and Social Media*.