SolomonLab at SemEval-2019 Task 8: Question Factuality and Answer Veracity Prediction in Community Forums

Sudeep Kumar Sahoo* **Ankita Gupta*** **Divya Prakash**
Rohit R.R **Vertika Srivastava** **Yeon Hyang Kim**

Samsung R&D Institute India, Bangalore
{gupta.ankita, sudeep.sahoo, p.divya,
rohit.r.r, v.srivastava, purine.kim}@samsung.com

Abstract

We describe our system for SemEval-2019, Task 8 on "Fact-Checking in Community Question Answering Forums (cQA)". cQA forums are very prevalent nowadays, as they provide an effective means for communities to share knowledge. Unfortunately, this shared information is not always factual and fact-verified. In this task, we aim to identify factual questions posted on cQA and verify the veracity of answers to these questions. Our approach relies on data augmentation and aggregates cues from several dimensions such as semantics, linguistics, syntax, writing style and evidence obtained from trusted external sources. In subtask A, our submission is ranked 3rd, with an accuracy of 83.14%. Our current best solution stands 1st on the leaderboard with 88% accuracy. In subtask B, our present solution is ranked 2nd, with 58.33% MAP score.

1 Introduction

With the rising popularity of online community question answering (cQA) systems such as Quora, StackOverflow, and Qatar Living forum (QLF), the amount of information shared over these platforms is also increasing rapidly with time. These forums provide effective information sharing mechanism to their users who can seek answers to their queries as well as post answers to the questions. However, the information shared on such platforms may not always be factual and correct. The responders may misunderstand the question being asked or merely ignore certain specific details. At times, the information shared may even be false or ambiguous in the desired context. This is aggravated by lack of moderation and systematic control on cQA forums. The Semeval-2019 Task 8[1] on "Fact Checking in Community Question Answering Forums" aims to solve this real-life problem.

The above task tries to explore the veracity of an answers to a question posted on QLF. While the precedent tasks such as SemEval (Nakov et al., 2015, 2016, 2017), address the issue of ranking answers according to their relevance to a question, the task-at-hand is the first one to consider the correctness of answers. This task is formulated as a two-stage problem. The first stage aims to identify the user posts asking for factual information. The answers to the identified factual questions are then fact-verified in the second stage. Both the subtasks are designed as 3-class supervised classification problems.

More specifically, the first stage or subtask A addresses the problem of determining whether the posted question asks for factual information, an opinion/advice or is just meant for socializing. For example, *"what is Ooredoo customer service number?"* asks for factual information, whereas *"What was your first car?"* is socializing and *"which is the best bank around?"* is seeking guidance/opinion. Each data sample in subtask A is a question posted by a user consisting of a subject, body and meta information (user ID, username, and the category of question, e.g., "Education," "Visa and Permits", "Welcome to Qatar" etc.).

The second stage or subtask B focuses on determining whether an answer to a factual question is true, false or does not constitute a proper answer, in which case, it is labeled as non-factual. For example, to the question *"Can I bring my pitbulls to Qatar?"*, Answer A1: *"Yes, you can bring it but be careful this kind of dog is very dangerous"* is factual-false[2], Answer A2: *"No, you cannot as they are banned"* is factual-true[2] and Answer A3: *"There goes another job opportunity for the sake*

*Equal Contribution

[1] http://alt.qcri.org/semeval2019/index.php?id=tasks

[2] can be verified at http://canvethospital.com/pet-relocation/

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1165–1171
Minneapolis, Minnesota, USA, June 6–7, 2019.

of two lovely animals. " is non-factual. The data is organized as a question-answer tuple: question posted by a user and an answer (body, username and answer ID) posted by the same or another user. It has been ensured that all the questions in this task are factual questions.

Our approach to solving this task is based on extracting rich-feature representation from the input and training a classifier to make predictions. The feature representation integrates knowledge from various complementary sources, such as the question/answer content, the content of other answers in the thread, evidence from trustworthy external sources of information, and the relevance of an answer to the question. For subtask A, we rely on question content (semantic, linguistic and syntactic cues), whereas the evidence from external sources and answer relevancy to the question are essential aspects for subtask B. For both the subtasks, we also leverage a data augmentation approach which facilitates the generalization ability of learned classifier on unseen test data as well as ameliorates the class imbalance issues present in the training data.

The rest of the paper is organized as follows: Section 2 gives an overview of our system. Section 3-5 describe the details of our approach. Section 6 demonstrates the experimental results. We conclude in Section 7.

2 System Overview

Our proposed system primarily relies on following key components (i) data augmentation (DA) (ii) pre-processing of question/answer content and (iii) feature extraction from multi-faceted sources.

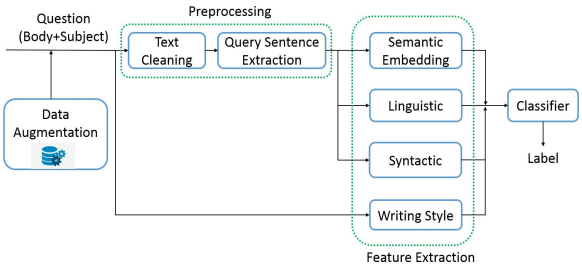

Figure 1: System Overview for Subtask A

As depicted in Figure 1, following DA and preprocessing of the question, our system for subtask A extracts semantic (what is said), linguistic (how it is said), syntactic (how it is structured) and writing style based features (how it is depicted) from the processed question. These extracted features are then combined to train a classifier for label prediction.

Subtask B also leverages DA and preprocessing as its first key steps. However, apart from features extracted for subtask A (as mentioned above), it also utilizes external evidence and forum-level features (Figure 2). The external evidence is collected from trusted sources using a search-engine. The forum-level features capture the relevancy of an answer to the question and its similarity to other answers in the same thread.

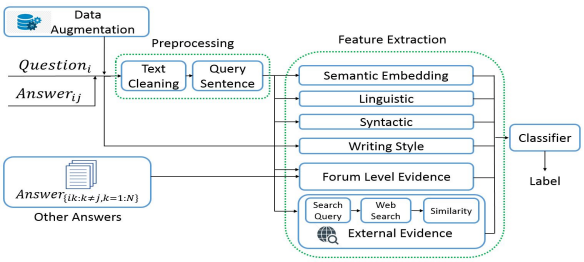

Figure 2: System Overview for Subtask B

3 Data Augmentation (DA)

Data augmentation (DA) is one of the main components of our proposed system that resulted in significant performance gains. For both the subtasks, the training data is imbalanced. This motivated us to look for ways to balance the distribution of data samples across classes and at the same time incorporate adversarial examples which are plausible in the real scenario but are not present in the training data. We next discuss the DA details for both the subtasks in the following subsections.

3.1 Subtask A

In the training data for subtask A, the number of samples from the "opinion" class (563) is observed to be twice as many samples from "factual" (311) or "socializing" class (244). In order to balance the class distribution, we sought to oversample both of the non-majority classes based on the domain knowledge.

For the "Factual" class, we leveraged the questions asked in subtask B. In subtask B, by its formulation, one is supposed to verify the veracity of answers to "factual questions." Thus, we used the training, development and test set of subtask B to augment training data for subtask A ("factual" class instances). This way, we extracted a total of 91 distinct factual questions. For the "socializing" class, we utilized the QL-unannotated-data[3] to se-

[3]additional resource by (Mihaylova et al., 2018)

Class	Question Body
Factual	**Can someone please tell me where can i find Garlic Oil in Qatar?** i heard it is good for hairfall. dont know if its true or not but really want to try it. So help me guys!
Opinion	**Is it right to resign from your job at this time of global crisis?** the reason is i'm not doing anything in the office. I feel useless; but I'm hesistant to resign because of the condition today even that I'm on husband sponsor.
Socializing	**Is this a beginning of a mutual friendship between Christianity and Islam in Qatar?** I hope they're going to sell some Bibles in Villagio coz I can't find somebody sellin' it around here.

Table 1: Example for query-sentence selection. The highlighted text is considered as the query-sentence.

lect samples from categories ("Funnies," "Good News Everyone," "Party on my mind," "Recipes," "Press Releases") that are assured to contain only socializing content. In these categories, the users are just trying to make conversation or share anecdotes. As the number of such samples is considerably large, we sample 320 samples (using reservoir sampling (Vitter, 1985)) to balance the distribution across classes in the original training data.

3.2 Subtask B

For subtask B, we consider an adversarial setting closely related to the problem at hand. As mentioned before, each data sample in this subtask is a question-answer tuple, and the answer can be either "true", "false" or "non-factual." A related task was demonstrated in Semeval 2016 task 3 "Answer Selection in cQA" (Nakov et al., 2016) where the objective was to re-rank the answers based on their relevancy to the question. In this task, the replies such as follow-up question from other user, clarifications, and acknowledgment from the user himself were categorized as "Bad" answers. Although, in the task-at-hand, the organizers have omitted such answers, in the real-life scenario they will also be present and should be categorized as "non-factual" in our current problem setting.

Thus, to include such samples, we extract factual questions from the training data of subtask A. For each of these questions, we select "bad answers" from the data provided in the SemEval 2016 task. The chosen question-answer pair is then annotated as "non-factual" and added to the training data of subtask B.

4 Preprocessing

Before feature extraction, we pre-process the input question/answer using several key steps. We expand the contractions and terms commonly used on social media platforms such as 'i'm: 'i am,' 'i'd: i would,' 'pls: please,' 'nt: not,' 'thru: through' etc. Furthermore, we use several markers such as URLs, images, emoticons, and punctuation marks in the question/answer to extract writing style and syntactic features (described in Section 5.3). For semantic and linguistic features, we strip these markers.

4.1 Query Sentence Extraction

Based on the empirical evidence, we could infer that the body of each question posed by the user contains several sentences. However, among all these sentences, only one or two convey the query he/she really wants to ask. Also, the user may post his question in the question subject itself. Thus, we extract these "query-sentences" from the question body and subject and use them to extract linguistic, semantic features. An example of the query-sentence and original question posted by the user for each of the three classes corresponding to subtask A is depicted in Table 1.

In order to extract query-sentence, we parse each sentence in the question using Stanford CoreNLP constituency parser (Manning et al., 2014). A sentence is considered a query-sentence if its parse-tree has *SBARQ*/*SQ* constituent phrases. We also use some common heuristics such as, whether the sentence ends with a question-mark or starts with common "wh" words (what, why, how etc).

5 Modeling Content : Feature Extraction

We use rich feature representation to model the information conveyed in question/answer. In the subsequent subsections, we describe the details of each of these features.

5.1 Semantic Sentence Embedding

Following the pre-processing step, we compute semantic sentence embedding for query-sentence by using two approaches. The first approach utilizes universal sentence encoder (USE) (Cer et al., 2018). It has been known to perform well with minimal amounts of supervised training data for a downstream task, which is precisely our setting for both the subtasks. The second approach appro-

priates pre-trained word embeddings (glove) (Pennington et al., 2014), averaged over each word in a sentence to compute sentence-level embedding.

5.2 Linguistic Features

Often, forum users exhibit linguistic cues in writing questions and answers. For example, they may use subjectively loaded words such as 'awesome,' 'worst' etc. while asking for an opinion rather than factual information. While answering on the forum, they may exhibit the degree of confidence in the truthfulness of what they say by using words like "most likely", "probably", "think" etc. We therefore use linguistic markers such as hedges (Hyland, 2018), weasels(Vincze, 2013), factives (Hooper, 1974), assertives (Hooper, 1974), implicatives (Karttunen, 1971), mood[4], modality[4], subjectivity[4], sentiment[4] and polarity of subjective words (Riloff and Wiebe, 2003) based on respective lexicons to compute a feature vector. (For details, refer to (Mihaylova et al., 2018))

5.3 Writing Style Features

We extract writing style features from the question/answer which capture the format of a user-post. A socializing question is more likely to be written informally as compared to factual/opinion query. A non-factual answer which is not much informative may also carry distinctive cues. To capture these aspects, we count the number of punctuations, emoticons, NON-ASCII characters and check the presence of URL, image, ALL CAPS, consecutive character repetition ($\geq$ 3 times). Table 2 depicts how the number of samples exhibiting a particular writing style feature vary across the three classes in subtask A. A similar trend is present for factual (true/false) versus non-factual answers in subtask B.

Feature	Opinion	Factual	Socializing
All Caps	8	1	4
URL	2	0	173
Image	0	0	3
Repetition	9	3	130

Table 2: Writing Style Based Feature Distribution across Classes for Subtask A

5.4 Syntactic Features

We also examine syntactic features such as part-of-speech (POS) and category of question encoded as bag-of-words features. Further, we consider the expected answer type for a question (QType) and named-entity-type (NET) in an answer.

QType suggests the kind of information the question is seeking such as "description", "entity", "human", "location", "number", "yes/no" and "others" (extracted using work in (Madabushi and Lee, 2016)). Such features help segregate the socializing class in subtask A. For subtask B; we exploit the relation between what type of information the user wants to ask (QType) and what type of information is provided in the answer (NET). To capture this, we extract the type of all named-entity mentions in the answer. We consider "person", "organization", "location" and "quantity" as possible NE tags extracted using spacy[5].

5.5 External Evidence

In subtask B, the verification of an answer requires external evidence to conclude about its veracity. We extract external evidence by formulating a search-query from the question and answer followed by a web search [6] of this search-query. For each of the obtained search results, we compute its similarity with the question and answer respectively. These similarity scores are then used as features to a classifier.

Search-Query Formulation In order to search the web for relevant evidence, we formulate a search-query based on the question and answer. We extract query-sentence from the question posted by a user and append "Qatar" if neither 'Doha' nor 'Qatar' is present in it.

Further, to incorporate relevant information from the answer into this search-query, we find the answer sentence that has the highest similarity with the query-sentence. From this top-ranked sentence, we extract up to 7 keywords based on named entities, noun-chunks[5] and unigrams sorted by tf-idf scores, where named entities and noun-chunks are given high priority. Query-sentence combined with keywords from the answer is used as search-query.

Search Results We collect search results (snippets) from reputed sources (e.g., news, government websites, official sites of companies) (Mihaylova et al., 2018) for search-query formulated as above. Since the search-query may not always be perfect, we also obtain search results by drop-

[4]https://www.clips.uantwerpen.be/pages/pattern-en

[5]https://spacy.io/usage/linguistic-features

[6]Bing web search API

Question	Answer	Evidence	Source	Class
how cold is doha during winter?	**8-10 degree** i guess	Over the course of the year, the temperature typically varies from **57 F (14 C) to 107 F (42C)** and is rarely below 51F or above 112F	Weatherspark	False
any private beaches in Qatar?	While Going to Shammal after 40 KM from Doha you will find **Al Ghariyah**	Qatar is a peninsula....This is a list of beaches in Qatar. Contents.**Al Ghariyah beach** is located 80 km north of Doha.	Wikipedia	True

Table 3: Example of external evidence collected in Subtask B

ping a few keywords from the search-query. From all the obtained search results, we select snippets that are most relevant to the question and the answer. Table 3 illustrates the external evidence retrieved for two question-answer pairs.

Similarity based Features For each question-answer pair, we compute their similarity with external search results obtained above. We use three similarity metrics: containment of unigrams, bigrams and trigrams (Lyon et al., 2001), cosine similarity of USE embedding and cosine similarity of tf-idf representation. For each metric, we compute the similarity of the snippet with: question, answer, query-sentence + top-answer sentence and all of them together. We then take the average and maximum over similarity scores for all the search results.

5.6 Forum-Level Features

These features capture the relevance of an answer with respect to a question as well as to other answers. An answer which contains information similar to that specified in other answers is more likely to be relevant and trustworthy. Thus, we consider the similarity of the answer with the question as well as its similarity with other answers in that thread. Here, also we consider all three similarity metrics mentioned before.

6 Experiments and Evaluation

6.1 Setting and Evaluation Metrics

We now utilize all features as portrayed in Figure 1 for subtask A and Figure 2 for subtask B. We train two separate SVM classifiers (Burges, 1998) on respective features for 3-class classification for both the subtasks. We use 10-fold cross validation for hyper-parameter tuning of SVM based on which, we choose "linear" kernel with C=0.5 (regularization parameter) for all the demonstrated experiments. All the results are reported on the test data with accuracy, recall, and F1 measure as evaluation metrics.

Additionally, we calculate Mean Average Precision (MAP) for subtask B, where the 'True' instances are considered relevant examples (in the context of Information Retrieval). MAP measures the capability of the system to predict 'True' instances with higher confidence.

6.2 Results for Subtask A

Table 4 shows the performance of the proposed system (PS). From the results, we can observe that our PS (excluding syntactic features) achieves an impressive performance with accuracy of 84.12% and 72.17% F1. Our submission (with all the features in PS + (POS and QType)) achieved similar performance and ranked 3rd on the leaderboard (83% acc., 71% F1) with only a marginal difference with respect to the first (84% acc., 72% F1) and second-ranked (83% acc., 72% F1) systems. To push our system's performance even further, we experimented in the post-evaluation phase and achieved 88.10% accuracy and 77.37% F1, highest on the post-evaluation leaderboard [7]. This current best solution leverages QType features, extensive data augmentation using bagging technique and excludes writing style features.

In order to appraise the importance of each feature, we conducted an ablation study by analyzing individual features and their combinations. It can be followed from the results that the semantic embedding contributes most to the performance of the system. However, embeddings derived using USE perform better than glove embeddings. This difference is possibly due to the failure of glove-based embedding in capturing word-order and long-range dependencies.

The second most important contributor is the data augmentation approach which resulted in notable accuracy gains (3.71% improvement). As expected, it allows the system to generalize better and ameliorates the issue of class imbalance. Following it is the query-sentence extraction approach with $\sim$1% accuracy enhancement. The

[7] As reported on 23/2/2019

Feature	Type	Acc	F1	Rec
PS+ Syntactic	POS	82.30	69.23	71.26
	Cat	83.77	71.61	73.52
	QType	84.67	73.00	74.75
PS	USE	**84.12**	**72.17**	**73.90**
	Glove	78.55	64.11	65.79
PS-WS		86.36	75.64	76.96
PS-Ling		83.49	72.09	74.94
PS-QSent		83.21	71.72	74.71
PS-DA		80.41	67.53	71.17
PS-Sem		70.97	52.95	54.92
Submission	Official	83.14	70.89	72.82
Best	Ensemble	**88.10**	**77.37**	**77.96**

Feature	Type	Acc	F1	Rec	MAP
PS+ Syntactic	POS				
	Cat				
	QType+NE	77.63	42.46	42.74	30.00
PS (Best)	USE	76.56	42.65	45.12	25.00
	Glove	**77.85**	**43.90**	**44.31**	**58.33**
PS-Ext		78.71	45.75	46.08	25.00
PS-Forum		75.48	41.30	42.25	62.50
PS-Sem		73.98	42.08	45.42	23.81
PS-Reputed		72.90	37.93	39.09	29.17
PS-DA		67.74	37.21	40.28	37.50
Submission	Official	69.00	37.44	40.25	33.33
Baseline	Majority	**83.01**	**28.47**	**33.33**	**15.55**

Table 4: Experimental Results. Subtask A (Left) and Subtask B(Right)
PS: Proposed System, WS: Writing Style, Ling: Linguistic, Qsent: Query Sentence, DA: Data Augmentation, Sem: Semantic Embedding, Cat: Category, Ext: External Evidence, Forum: Forum level evidence, NE: Named Entity from answer

performance is in line with the expectations as query-sentences are sufficient in capturing the essence of user question. QType and linguistic cues help improve the performance further.

However, we notice that the performance improves by excluding the writing style features. The possible reason for this observation can be the absence of such features in the test data. In the training and development data-sets, the presence/absence of these features was a distinguishing factor among classes (see Table 2) which made them worth considering.

6.3 Results for Subtask B

Table 4 shows the performance of PS with the ablation study. Our PS (also our best[7]) achieves an accuracy of 77.85% with 58.33 MAP (highest among all the participating systems). From the ablation study, we observe that although removal of external evidence results in slight accuracy gains (0.86%), it causes a drastic reduction in MAP score (33.33 points). This signifies the importance of external evidence as these features enable the system to make better predictions for the true/false classes.

We also conducted a majority baseline experiment where all the samples are labeled as "non-factual." This experiment resulted in the best performance on the leaderboard with 83% accuracy and very poor MAP. This illustrates that the test data has a majority of non-factual instances. Thus, measuring the performance of any system solely on accuracy for this problem is not fair.

As it can be inferred from the ablation study, among all the features, reputed source based search-results selection (contributing 4.95% acc. gain) and forum-level features (contributing 2.37% acc. gain) are the most important. Reputed source selection helps in relying on only trusted sources for external evidence selection and hence make better predictions for true/false classes. Forum level features help in distinguishing among non-factual versus true/false samples.

Further, data augmentation in subtask B results in significant performance gains of 10.11% accuracy. It helps the system learn about the characteristics of "bad answers" which are not present in the training data and hence enables the system to generalize better on the test data. For this subtask as well, semantic embedding plays a vital role in capturing the essence of the question-answer pair, contributing 3.87% gain in accuracy.

7 Conclusion

In this work, we have described our system for Semeval-2019 Task 8 on Fact-checking in cQA Forums. Our system leverages data augmentation and integrates knowledge from various aspects, such as the semantics, linguistics, syntax and writing style along with complementary information from trustworthy external sources and QLF.

Our submission was ranked third in Subtask A with marginal performance differences compared to the best-ranked systems. Our current best solution is ranked first on the leaderboard with 88% accuracy[7]. In subtask B, our current best solution is ranked 2nd, with 58.33% MAP score, highest among all participating systems[7].

However, none of the participating systems could beat the majority baseline for subtask B in terms of accuracy, which signifies that we are still far from solving this task to its entirety with a decent performance. Thus, there remains a lot of potential in this research direction.

References

Christopher JC Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Joan B Hooper. 1974. *On assertive predicates*. Indiana University Linguistics Club.

Ken Hyland. 2018. *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

Lauri Karttunen. 1971. Implicative verbs. *Language*, pages 340–358.

Caroline Lyon, James Malcolm, and Bob Dickerson. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the*.

Harish Tayyar Madabushi and Mark Lee. 2016. High accuracy rule-based question classification using question syntax and semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1220–1230.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tsvetomila Mihaylova, Preslav Nakov, Lluis Marquez, Alberto Barron-Cedeno, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. Semeval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 269–281.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*.

Veronika Vincze. 2013. Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles.

Jeffrey S Vitter. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57.