Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing

Noushin Rezapour Asheghi, Serge Sharoff, Katja Markert Emails: {scs5nra, s.sharoff, scskm }@leeds.ac.uk

UNIVERSITY OF LEEDS

Motivation

Several efforts have been made to build genre annotated web corpora and to employ them for research in the field of automatic genre identification. The major shortcomings of existing genre-annotated web corpora are:

- **Reliability**: Low inter-coder agreement
- **Size**: not large enough to ensure representativeness of genre classes
- **Format**: preserved in different formats such as PDF or plain text which results in losing HTML tags
- **Topic Diversity**: collected from a small number of sources which are topically similar

Challenges

- There is no universally agreed set of genre labels.
- There is disagreement in definitions, boundaries and granularities of genre labels.

Building a Reliable Genre-annotated Corpus

The genre web corpus should fulfil the following criteria:

- It needs to be reliable.
- It must be collected from a diverse range of sources in order to avoid creating false correlations between genres and topics.
- It must include genre classes which are exclusive to the web.
- Web pages must be saved in HTML format.

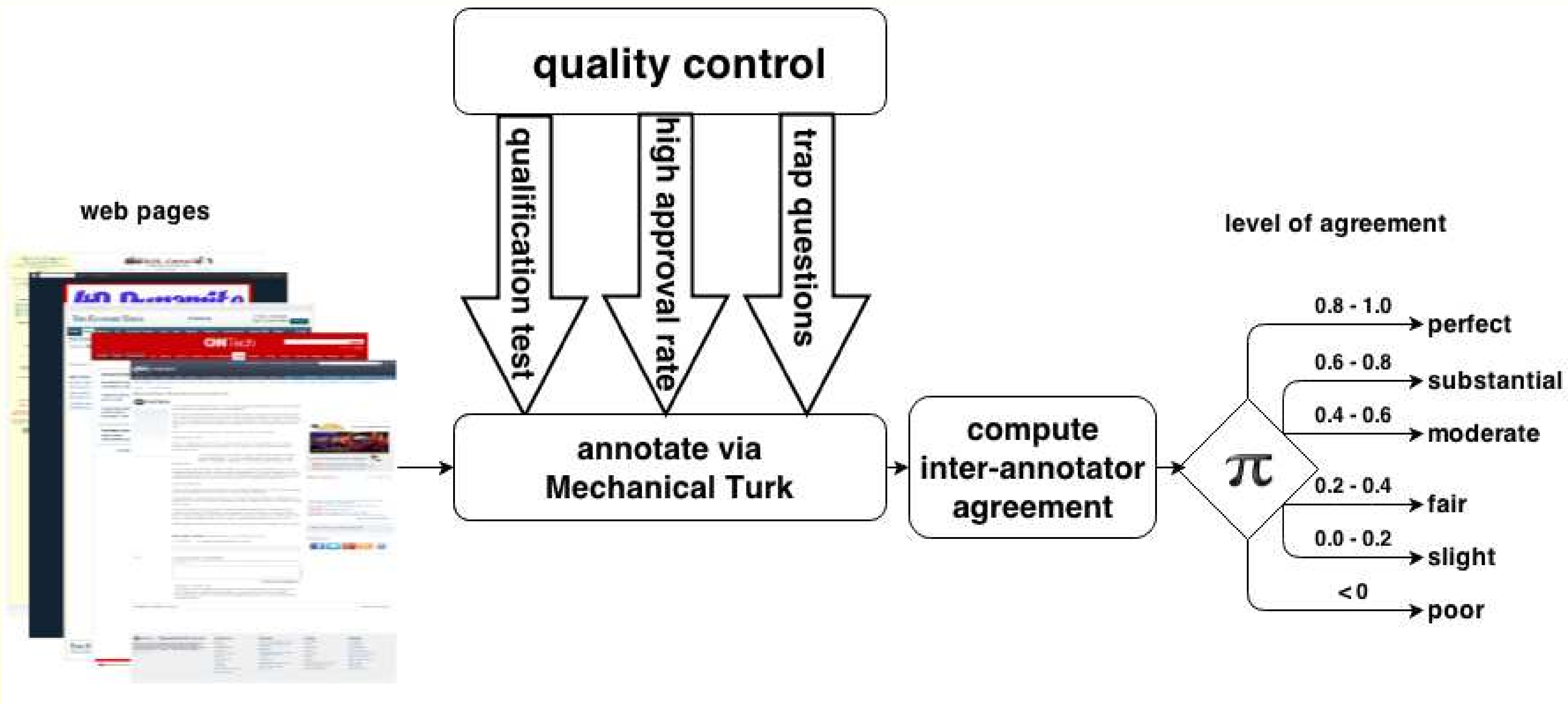
Corpus Compilation

Types of web corpora:

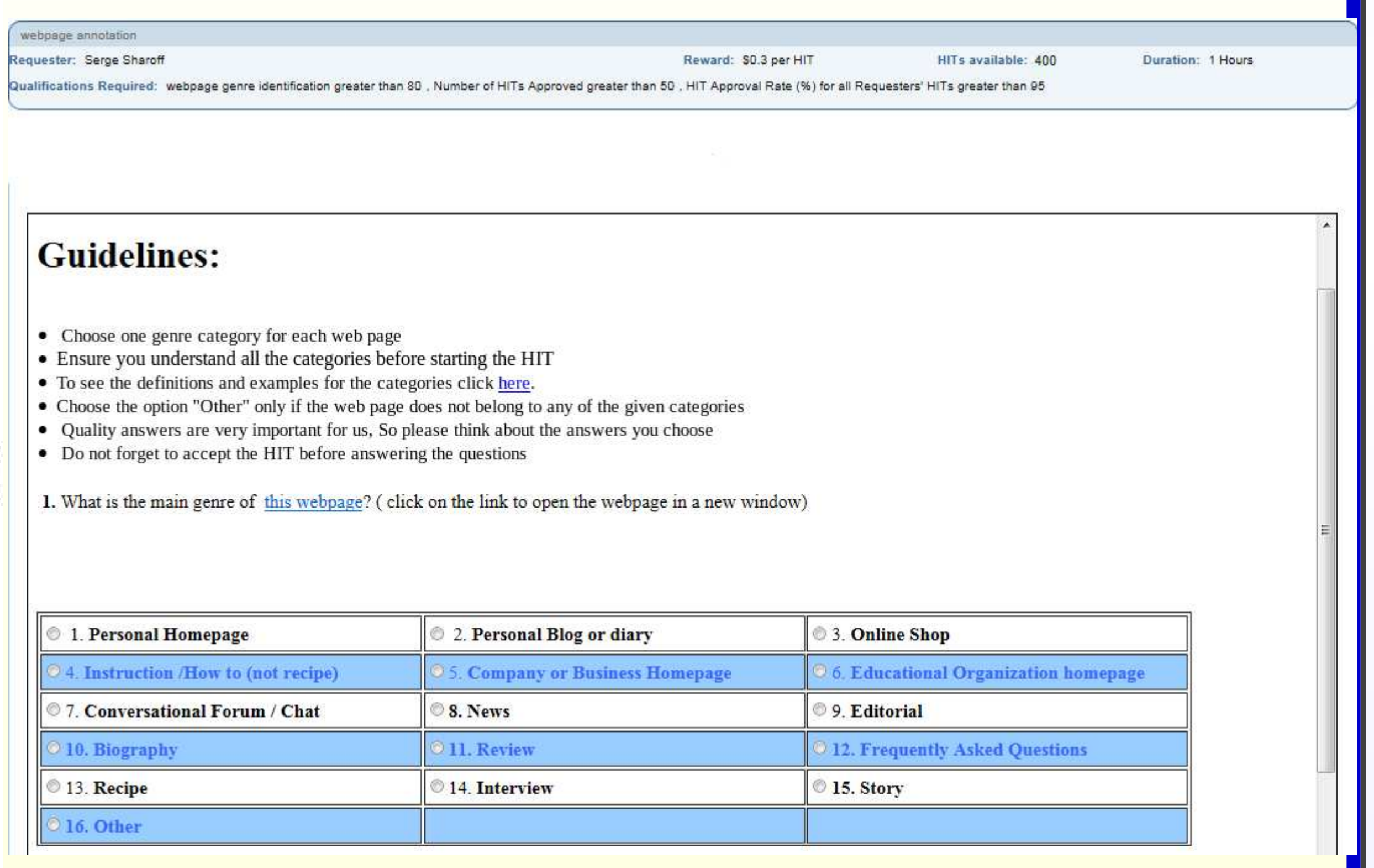
- **Designed corpus**: Time consuming; we have control on the content of a corpus (select from various topics)
- **Crawled corpus**: Fast; Less control on the content of a corpus; No guarantee to get a balanced corpus with a large number of web pages for each category

We chose to build a designed corpus because we wanted a balanced and topic diverse corpus.

Corpus Annotation via Amazon Mechanical Turk



In order to ensure high quality annotation, we restricted the range of workers who can complete our task. We only allowed workers who had completed at least fifty previously accepted HITs; have approval rate higher than 95% and pass our qualification test with the score of equal or higher than 80%.



webpage annotation

Requester: Serge Sharoff Reward: $0.3 per HIT HITs available: 400 Duration: 1 Hours

Qualifications Required: webpage genre identification greater than 80 , Number of HITs Approved greater than 50 , HIT Approval Rate (%) for all Requesters' HITs greater than 95

Guidelines:

- Choose one genre category for each web page
- Ensure you understand all the categories before starting the HIT
- To see the definitions and examples for the categories click here.
- Choose the option "Other" only if the web page does not belong to any of the given categories
- Quality answers are very important for us, So please think about the answers you choose
- Do not forget to accept the HIT before answering the questions

1. What is the main genre of this webpage? (click on the link to open the webpage in a new window)

○ 1. **Personal Homepage**	○ 2. **Personal Blog or diary**	○ 3. **Online Shop**
○ 4. **Instruction /How to (not recipe)**	○ 5. **Company or Business Homepage**	○ 6. **Educational Organization homepage**
○ 7. **Conversational Forum / Chat**	○ **8. News**	○ 9. **Editorial**
○ **10. Biography**	○ **11. Review**	○ **12. Frequently Asked Questions**
○ 13. **Recipe**	○ 14. **Interview**	○ **15. Story**
○ **16. Other**		

This is the screen-shot of genre annotation task on Mturk website. One of the defined genre labels in the guidelines or the option “other” can be chosen for each web page.

Results of the Annotation Study

Genre	Number of		# of pages from the same website			π
	web pages	websites	max	min	med	
Personal Homepage	304	288	9	1	1	0.858
Company/ Business Homepage	264	264	1	1	1	0.713
Educational Organization Homepage	299	299	1	1	1	0.953
Personal Blog /Diary	244	215	9	1	1	0.812
Online Shop	292	209	23	1	1	0.830
Instruction/ How to	231	142	15	1	1	0.871
Recipe	332	116	8	1	1	0.971
news	330	127	12	1	1	0.801
Editorial	310	69	11	1	3	0.877
Conversational Forum	280	106	11	1	1	0.951
Biography	242	190	15	1	1	0.905
Frequently Asked Questions	201	140	8	1	1	0.915
Review	266	179	15	1	1	0.880
Story	184	24	38	1	7	0.953
Interview	185	154	11	1	1	0.905

Statistics for each category illustrate source diversity and reliability of the corpus.

Conclusions

We present the first web genre corpus which is reliably annotated. we used crowd sourcing which is a novel approach in genre annotation. The result of inter-coder agreement shows that the corpus has been annotated reliably. Table below gives an overview of the corpus statistics.

Number of genres	15
Number of web pages	3964
Number of web pages for the smallest category	184
Number of web pages for the largest category	332
Median Number of web pages for the categories	266
Number of tokens	7,205,820
Number of types	130,254
Number of sentences	329,861

The corpus statistics

The future work involves extending this corpus by using random web pages. We also plan to extend the number of genre classes.