Crowdsourcing Annotation
for Machine Learning
in Natural Language Processing Tasks

(NON-FINAL VERSION!
Proofread version will be uploaded April 30, 2012.)

by

Omar F. Zaidan

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland
April, 2012

© Omar F. Zaidan 2012
All rights reserved
Abstract

Human annotators are critical for creating the necessary datasets to train statistical learners, but annotation cost and limited access to qualified annotators forms a data bottleneck. In recent years, researchers have investigated overcoming this obstacle using crowdsourcing, which is the delegation of a particular task to a large group of untrained individuals rather than a select trained few.

This thesis is concerned with crowdsourcing annotation across a variety of natural language processing tasks. The tasks reflect a spectrum of annotation complexity, from simple labeling to translating entire sentences. The presented work involves new types of annotators, new types of tasks, new types of data, and new types of algorithms that can handle such data.

The first part of the thesis deals with two text classification tasks. The first is the identification of dialectal Arabic sentences. We use crowdsourcing to create a large annotated dataset of Arabic sentences, which is used to train and evaluate language models for each Arabic variety. We also introduce a new type of annotations we call annotator rationales, which complement traditional class labels. We collect rationales for dialect identification and for a sentiment analysis task on movie reviews. In both tasks, adding rationales yields significant accuracy improvements.

In the second part, we examine how crowdsourcing can be beneficial to machine translation (MT). We start with the evaluation of MT systems, and show the potential of crowdsourcing to edit MT output. We also present a new MT evaluation metric, RYPT, that is based on human judgment, and well-suited for a crowdsourced setting. Finally, we demonstrate that crowdsourcing can be used to collect translations. We
ABSTRACT

discuss a set of features that help distinguish well-formed translations from those that are not, and show that crowdsourcing yields high-quality translations at a fraction of the cost of hiring professionals.

Thesis Committee:

Chris Callison-Burch
Associate Research Professor
Department of Computer Science
Johns Hopkins University

David Yarowsky
Professor
Department of Computer Science
Johns Hopkins University

Christine D. Piatko
Principal Research Scientist
Johns Hopkins University Applied Physics Laboratory
Acknowledgments

My awesome advisor told me that a majority of people reading a Ph.D. thesis “will only ever read the Introduction and Conclusion.” If it were up to me, I’d want them to read this part too, because those mentioned below are why this thesis exists.

First and foremost, I thank my family. Mom and Dad, thank you for the HUGE amounts of love and support. Ahmad, you have no idea how much I have learned from you. You are a freaking genius. Mona, thank you for showing me that two people who might seem different are actually so similar. Also, thank you for the brilliant music. And thank you guys for bringing Gergana and Na’el into our lives. Maya, Hanin, Shams, and Adam make me extremely thankful to be an uncle.

And then there’s Khaled, one of the smartest and most amazing guys I know. If my ‘mentorship’ had anything to do with that, then I’m incredibly proud. Thank you for the memories. You were once my adorable baby brother (seeboo!), but became a person for whom I have much respect, in whom I place my trust, and with whom I want to do everything. I only wish I could spend even more of my time with you.

If there’s one person who suffered with me through my Ph.D., that would be my Dima. Boo, thank you for toughing it out with me. I can’t wait to meet our kids and tell them our story. Thank you most importantly for showing me the other way of doing things. I love you, and we are going to do great things together. I promise.

My thesis advisor, Chris Callison-Burch, played a big part in shaping and promoting my research. He helped with too many things to even try to recount. I’m also immensely proud of our work, and cherish the fact that I have a CCB number of 1. And on a personal level, he is the most affable, supportive, and just plain fun person.

1I forgive you for the LEGO. You forgive me for the unfinished airport, yes?
ACKNOWLEDGMENTS

Chris, I miss hanging out with you already. I wish you and Dawn all the best.

Christian Scheideler was my first advisor at JHU, but by my 2nd year, had accepted a position at TU München. Christian helped me survive a tough first semester, and I also say Danke vielmals for the offer to accompany him to Munich. Jason Eisner then played a pivotal role advising me, and our collaboration resulted in several papers and got my foot in CLSP’s door. Jason, thank you for being an amazing teacher, and for taking a chance on a grad student whose only exposure to NLP was your course.

David Yarowsky and Christine Piatko rounded out my thesis committee, and provided many suggestions to improve the quality of this document.\footnote{Naturally, all remaining errors are either mine, or the doing of a very sneaky troll.} I also want to thank my GBO committee: Chris, Jason, Christine, Sanjeev Khudanpur, and Mark Dredze. I was ten times more nervous about the GBO than my defense(!), but you guys turned it into an enjoyable conversation about my work.

CS and CLSP staff greatly reduced the amount of stress I endured. At CS, Debbie and Cathy have helped with scheduling and (a lot of) paperwork. At CLSP, Monique, Desirée, and Justin have helped me with payroll, more paperwork, and IT support. Much thanks goes to Laura Graham, who I’m convinced basically knows everything, and Steve Rifkin, who is always super responsive and super friendly.

I’ve had fantastic educators as an undergrad at SLU. Thank you Mike Sheard, Jeff Greathouse, and the folks at MCSS (Maegan Bos, Jim DeFranza, Ed Harcourt, Brian Ladd, Patti Frazer Lock). In my formative years, Ms. Lamia Shloudi taught me how to think mathematically, and Mr. Salah Rammal taught me BASIC and made me realize, hey, I can make a computer do useful stuff. Ms. Fatima Dabbas and Mr. Omar Tayeh launched me on a great career path. Thank you all. Also, thank you Ms. Rula Kamal, for being, simply, the best.

Whether they know it or not, I learned a lot from: Eftekhar Al-Farkh (RIP), Laila Bustami, Rami Bustami (RIP), Sam Carliles, Rance Davis, Nizar Habash, Kenneth Heafield, John Jaunzems, Fred Jelinek (RIP), Zhifei Li, John Makhoul, Ben Mitchell, and Noah Smith. I thank you!

I have been fortunate to have made really awesome friends at JHU/Baltimore:
ACKNOWLEDGMENTS

Abby, Alex “Sasha” Klementiev, Charley, Hassan, Jay, Mahmoud, Megan, Mike L., Osama, Sammy, and Siddharth A. Thank you for the great and generally hilarious times. To my pre-JHU dear friends Alba, Alejandro, Hamza, Mais, and Mohammad A.S.: thank you for pretending to be so interested in what I do. To Abdul, Ammar, Amr, Bashar I./S./T., Lara, and Lindsay: one of my greatest regrets is not being able to stay in touch with you as well as I wanted. I’m sorry for that.

Last but not least, Juri Ganitkevitch is all kinds of awesome. He is a bauce, my friend, my bro, and, yes, my dawg. YOU NEED TO BEFRIEND HIM IMMEDIATELY.


During my time at JHU, I was supported by several teaching assistanships from the CS department, as well as DARPA’s GALE Project, the WSE-APL Partnership Fund, and Raytheon BBN Technologies. Further funds to support my work also came through EuroMatrix, NSF, Microsoft, and Google.

Thank you JHU for inviting many great speakers. Thank you Microsoft and Google for developing great technology. Thank you Erin Currier for the portrait on the next page.³ Thank you Zaidans for a 17/12 that will go down in history. And thank you Chipotle, Billy Joel, and Jon Stewart for helping me maintain my sanity.

³URL: http://erincurrierfineart.com/mBouazizi.html
This thesis is dedicated to
Mohammed Bouazizi
(29 March 1984 - 4 January 2011)
Contents

Abstract ii

Acknowledgments iv

List of Tables xiii

List of Figures xv

1 Introduction 1

1.1 Major Themes and Overview . . . . . . . . . . . . . . . . . . . . . . . 2

1.1.1 Quality Control . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.1.2 Novel Annotation Schemes (and Novel Models Benefiting from
Them) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.1.3 Reducing Annotation Cost . . . . . . . . . . . . . . . . . . . . . 7

1.1.4 Access to Linguistically Qualified
Annotators . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

1.2 Main Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

1.3 Thesis Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

1.4 Publications Resulting from this Thesis . . . . . . . . . . . . . . . . 13

2 Crowdsourcing and Amazon’s Mechanical Turk 15

2.1 Amazon’s Mechanical Turk Service . . . . . . . . . . . . . . . . . . 16

2.1.1 The Demographics of Mechanical Turk . . . . . . . . . . . . . 18
CONTENTS

2.2 Literature Review .................................................. 19
  2.2.1 Annotation for NLP Tasks on MTurk ......................... 19
  2.2.2 Quality Control on MTurk .................................... 21
2.3 MTurk Best Practices .............................................. 22
2.4 Typical Turker Behavior and MTurk Sampling .................... 25
2.5 Conclusion ........................................................... 29

3 Arabic Dialect Identification ....................................... 30
  3.1 Background: The MSA/Dialect
    Distinction in Arabic .............................................. 31
    3.1.1 The Dialectal Varieties of Arabic .......................... 33
    3.1.2 Existing Arabic Data Sources .............................. 36
    3.1.3 The Arabic Online Commentary Dataset .................... 37
  3.2 Arabic Dialect Identification ................................. 39
    3.2.1 The Difficulty of Arabic DID ............................... 40
    3.2.2 Applications of Dialect Identification .................. 43
  3.3 Crowdsourcing Arabic Dialect Annotation ..................... 44
    3.3.1 Annotation Interface ....................................... 44
    3.3.2 Annotator Behavior ....................................... 45
      3.3.2.1 Label Distribution ................................... 48
      3.3.2.2 Annotator Agreement and Performance ............... 50
      3.3.2.3 Annotator Bias Types ................................ 51
  3.4 Models for Automatic Dialect Identification .................. 54
    3.4.1 Smoothed N-Gram Models .................................. 54
    3.4.2 Baselines ................................................ 55
    3.4.3 Experimental Results .................................... 55
      3.4.3.1 MSA vs. Dialect Classification ....................... 55
  3.5 Related Work ................................................... 63
  3.6 Conclusion ...................................................... 64

ix
## CONTENTS

### 4 Annotator Rationales for Text Classification

- 4.1 Annotator Rationales
  - 4.1.1 Why Would Rationales Help?
  - 4.1.2 The Movie Review Polarity Dataset
  - 4.1.3 Data Collection
  - 4.1.4 Notation and Features
- 4.2 Integrating Rationales into a Discriminative Model
  - 4.2.1 Contrast Examples
- 4.3 Integrating Rationales into a Generative Model
  - 4.3.1 Modeling Rationales Explicitly
  - 4.3.2 \( p_{\phi} \) as a Conditional Random Field
    - 4.3.2.1 \( p_{\phi} \)'s Emission Features
    - 4.3.2.2 \( p_{\phi} \)'s Transition Features
- 4.4 Experimental Results
  - 4.4.1 Experimental Design
  - 4.4.2 Tuning the Modified SVM's Hyperparameters
  - 4.4.3 Optimization of the Generative Model's \( \vec{\theta} \) and \( \vec{\phi} \)
  - 4.4.4 Results
- 4.5 Case Study II: Dialect Identification
  - 4.5.1 Results
- 4.6 Cost Analysis
- 4.7 Related Work
- 4.8 Conclusion

### 5 Crowdsourcing Manual Evaluation of Machine Translation Systems

- 5.1 Background: MT Evaluation
  - 5.1.1 Automatic Evaluation Metrics
    - 5.1.1.1 The BLEU Metric
    - 5.1.1.2 The TER Metric
  - 5.1.2 Criticisms of Fully-Automatic Scoring
  - 5.1.3 Crowdsourcing Manual Evaluation
CONTENTS

5.2 The HTER Metric ........................................... 107
  5.2.1 Crowdsourcing Editing ............................... 108
  5.2.2 Datasets ............................................... 110
  5.2.3 Turkers’ Editing Behavior ........................... 112
  5.2.4 Experiments ........................................... 113
    5.2.4.1 HTER Predictors ................................. 113
    5.2.4.2 Editor Calibration ............................... 117
5.3 The RYPT Metric .......................................... 120
  5.3.1 Obtaining Source-to-Candidate Alignments .......... 122
  5.3.2 Crowdsourcing RYPT Judgments ..................... 123
  5.3.3 Label Percolation .................................... 127
5.4 Human-Based Parameter Tuning of MT Systems .......... 130
  5.4.1 Minimum Error Rate Training ....................... 131
  5.4.2 Och’s Line Search Method .......................... 132
  5.4.3 Building a Human Judgment Database ............... 134
  5.4.4 Evaluating RYPT as an MT Metric .................. 136
5.5 Related Work ............................................. 139
5.6 Conclusion ............................................... 140

6 Crowdsourcing Translation .................................. 142
  6.1 Creating a Parallel Dataset ............................ 143
    6.1.1 Translation by Non-Professionals ................. 144
  6.2 Data Collection ......................................... 147
    6.2.1 Translation HIT Design ............................ 147
    6.2.2 Post-editing and Ranking HITs ..................... 149
    6.2.3 Data Collection Cost .............................. 151
  6.3 A Selection Model for Quality Control ............... 151
    6.3.1 Model Features .................................... 152
    6.3.2 Parameter Tuning .................................. 154
    6.3.3 The Worker Calibration Feature ................. 154
CONTENTS

6.4 Experimental Results ............................................. 155
  6.4.1 Evaluation Strategies ....................................... 155
  6.4.2 Translation Quality: BLEU Scores Against Professionals ... 156
  6.4.3 Fitness for a Task: Ranking MT Systems ................. 158
  6.4.4 Analysis .................................................... 158
6.5 Crowdsourcing Translation of Dialectal Arabic ............... 162
6.6 Related Work .................................................. 163
6.7 Conclusion ..................................................... 166

7 Conclusion .......................................................... 168
  7.1 Major Contributions ........................................... 169
  7.2 The Future of Mechanical Turk ............................... 170
  7.3 Future Work .................................................. 173

A The Buckwalter Transliteration Scheme ........................... 176

Vita ................................................................. 194
## List of Tables

3.1 A few examples illustrating the differences across MSA and three Arabic dialects. ............................................. 35
3.2 A summary of the different components of the Arabic Online Commentary dataset. ............................................. 39
3.3 Some statistics over the labels provided by three spammers. .......... 47
3.4 The specific-dialect label distribution (given that a dialect label was provided), shown for each speaker group. .................. 53
3.5 Two annotators with a General label bias, one who uses the label liberally, and one who is more conservative. .......... 53
3.6 Accuracy rates on several classification tasks for various models. . 56
4.1 Accuracy rates using each annotator’s data. .......................... 90
4.2 Accuracy rate for an annotator’s $\theta$ obtained when using some other annotator’s $\phi$. ................................. 91
4.3 Cross-entropy per tag of rationale annotations $\vec{r}$ for each annotator, when predicted from that annotator’s $\vec{x}$ and $\vec{\theta}$ via a possibly different annotator’s $\phi$. ............................... 91
4.4 An annotation cost comparison between the standard SVM and the modified SVM. ................................. 97
5.1 Summary statistics for each genre in the dataset. ....................... 112
5.2 Document ranking correlation for the different HTER-predictors, across the four genres. ................................. 117
5.3 The label distribution for collected judgments, for each source substring length. ............................................. 127
5.4 Results of the two RYPT vs. BLEU comparison experiments. ....... 137
5.5 Results of the RYPT vs. BLEU comparison experiment, grouped by sentence. ............................................. 138
6.1 LDC’s less commonly taught languages and their speaker counts. ... 145
LIST OF TABLES

6.2 The ability of different selection methods to reproduce a BLEU ranking of 6 MT systems obtained using professional translations as references. 159
6.3 BLEU scores and OOV rates for two dialectal test sets, one Egyptian and one Levantine, using different training corpora. . . . . . . . . . . . 165
List of Figures

1.1 Several translations for an Urdu sentence, produced by professional and non-professional translators. 3
1.2 The effect of quality control on crowdsourced translation quality. 4
1.3 A movie review that has had textual segments of it highlighted as rationales supporting a negative class label. 5
1.4 An illustration of the modified SVM of Chapter 4. 6
1.5 A map showing Middle Eastern countries, where various dialectal Arabic varieties are spoken. 10

2.1 MTurk’s HIT count over the first three months of 2012. 17
2.2 The percentage of the approved data contributed by the x most prolific Turkers. 26
2.3 The number of assignments completed from all HITs and the portion of assignments completed from completed HITs only, during the first 50 days of a data collection effort. 27
2.4 The sampling order of the HITs during the same data collection effort of Figure 2.3. 28

3.1 One possible breakdown of spoken Arabic into dialect groups. 33
3.2 Two roughly equivalent Arabic sentences, one in MSA and one in Levantine Arabic, translated by the same MT system into English. 37
3.3 The output of a Spanish-to-English system when given a Portuguese sentence as input, compared to the output of a Portuguese-to-English system. 38
3.4 Three heavily dialectal sentences that do not contain individually dialectal words. 41
3.5 The dialectal sentences of Figure 3.4, with MSA equivalents. 42
3.6 The interface for the dialect identification task. 46
LIST OF FIGURES

3.7 The distribution of labels provided by the workers for the dialect identification task, over all three news sources, and over each individual news source. ................................................................. 49
3.8 A bubble chart showing workers’ MSA and dialect recall. .......... 51
3.9 Learning curves for the general MSA vs. dialect task, with all three news sources pooled together. ...................................................... 57
3.10 Learning curves for the MSA vs. dialect task, for each of the three news sources. ................................................................. 58
3.11 Accuracy rate vs. sentence length. ........................................... 59
3.12 Words with the highest and lowest dialectness factor values. .... 61
3.13 A plot of the most common words in the $Al-Ghad$ sentences, showing each word’s $DF$ and corpus frequency. ................................ 62
3.14 A plot of the most common letters in the $Al-Ghad$ sentences, showing each letter’s $DF$ and corpus frequency. .......................... 62
4.1 An example of a negative review. ............................................. 69
4.2 The example review from Figure 4.1, translated into Arabic, and annotated with rationales that support a negative class label. .... 70
4.3 The example review from Figure 4.1, annotated with rationales that support a negative class label. ............................................. 71
4.4 Histograms of rationale counts per document. .......................... 73
4.5 An illustration comparing the standard SVM to the modified SVM. 78
4.6 Modeling rationales as sequence annotation. ............................ 80
4.7 The function family $B_s$ of Equation 4.13, for several values of $s$. 82
4.8 Classification accuracy curves for two baseline learners that only use class data, and two learners that also utilize rationale annotations. . . 89
4.9 Examples of Arabic sentences, with dialectal portions highlighted. 92
4.10 A precision-recall scatter plot for Turkers working on the dialect rationale annotation task. ...................................................... 94
4.11 Learning curves for the MSA vs. dialect classification task for the methods presented in this Chapter. ......................................... 95
5.1 An Urdu source sentence with multiple correct English translations. 102
5.2 Fluency and adequacy scales in the LDC specification. ............... 103
5.3 BLEU’s brevity penalty as a function of candidate length and reference length. ................................................................. 105
5.4 An example of minimally editing an MT output to match a pre-existing reference in meaning. ...................................................... 107
5.5 A scatter plot of documents’ TER and BLEU scores vs. HTER score. 109
5.6 The interface for the editing task. .............................................. 111
5.7 Scatter plots of Turker’s edit rate (on the MT output) vs. the rate required to produce the NIST reference from their submissions. . . 114
LIST OF FIGURES

5.8 Scatter plots of Turker’s edit rate (on the MT output) vs. the rate of the LDC editor (on the same MT output) .................. 115
5.9 Document ranking correlation for the calibration approach, plotted against varying sizes of the calibration set .................. 119
5.10 A source sentence and its parse tree, aligned with a candidate translation and its derivation tree .................. 121
5.11 An example of resolving a phrasal alignment to produce a word alignment ........................................... 124
5.12 The interface for collecting acceptability judgments .................. 126
5.13 Label percolation under different maxlen values .................. 130
5.14 Och’s method applied to a set of two foreign sentences .................. 133

6.1 Several translations for an Urdu headline, produced by professional and non-professional translators .................. 146
6.2 The interface for the translation task .................. 148
6.3 Redundant translations for several source sentences, solicited from different Turkers .................. 150
6.4 BLEU scores for different selection methods, measured against the reference sets .................. 157
6.5 The effect of varying the amount of calibration data (and using only the calibration feature) .................. 160
6.6 BLEU scores for the five right-most setups from Figure 6.4, constrained over the original translations .................. 161
6.7 Four dialectal Arabic sentences translated into English by two different MT systems, one trained on MSA-only data, and the other trained on crowdsourced dialectal Arabic data .................. 164

A.1 The ASCII-to-Arabic mapping used in Buckwalter transliteration ........................ 177