Low-Quality Training Data in Information Extraction

Diego Marcheggiani[1] and Fabrizio Sebastiani[2]

[1]Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy
[2]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar

Introduction

- Little or no prior work on investigating the effects of the quality of training data on IE accuracy
- Low quality of training data may have different causes:
  - Cost issues may have been more important than quality issues at annotation time;
  - The coders entrusted with the work may not have been involved in the design of the concept set;
  - The training data may be outdated.
- Common denominator among the above is that
  - a non-authoritative coder C_β annotated Tr;
  - an authoritative coder C_α (defined as the one who annotated Te) would have annotated Tr differently.

Methodology

- Goal of our work: measuring how much accuracy suffers when Tr is annotated by C_β.
- We do this by comparing:
  - accuracy in an authoritative setting (i.e., both Tr and Te annotated by C_α).
  - accuracy in a non-authoritative setting (i.e., Te annotated by C_α and Tr annotated by C_β).
- We use the token-and-separator F_1 measure to evaluate annotation accuracy, and Cohen's kappa (κ) to evaluate intercoder agreement.

Dataset

- We perform experiments, using LC-CRFs and HM-SVMs as learners, on UmbertoI(RadRep), a clinical IE dataset of 500 mammographic reports written in Italian and annotated according to 9 concepts (e.g., FollowupTherapies, OutcomesOfSurgery, etc.).
- The reports were annotated by 2 equally expert radiologists:
  - 191 reports by Coder1 only (`1-only`)
  - 190 reports by Coder2 only (`2-only`)
  - 119 reports by Coder1 and Coder2 (`Both(1)` and `Both(2)`)

	Mentions	Tokens
Annotated by Coder1	1,045	18,529
Annotated by Coder2	1,210	24,822

Experimental Protocol

- Two batches of experiments:
  - Batch1: Coder1 is C_α, i.e., Te is `1-only`. Tr is `Both(1)` in the authoritative setting and `Both(2)` in the non-authoritative setting.
  - Batch2: Coder2 is C_α, i.e., Te is `2-only`. Tr is `Both(2)` in the authoritative setting and `Both(1)` in the non-authoritative setting.
- We also test partially authoritative settings, i.e., a randomly chosen $\lambda\%$ of Tr is annotated by C_β, and the rest is annotated by C_α.

Extraction accuracy for the two settings

			LC-CRFs		HM-SVMs	
	λ	$\kappa(\lambda)$	F_1^μ	F_1^M	F_1^μ	F_1^M
Batch1	0	1.000	0.783	0.674	0.820	0.693
	100	0.742	0.765 (-2.35%)	0.668 (-0.90%)	0.786 (-4.33%)	0.688 (-0.73%)
Batch2	0	1.000	0.808	0.752	0.817	0.754
	100	0.742	0.733 (-10.23%)	0.654 (-14.98%)	0.733 (-11.46%)	0.625 (-20.64%)
Average	0	1.000	0.795	0.713	0.819	0.724
	100	0.742	0.749 (-6.14%)	0.661 (-7.87%)	0.760 (-7.76%)	0.657 (-10.20%)

Main Findings

- F_1 as a function of λ varies much less for Batch1 than for Batch2.
- We conjecture this to be due to the fact that Coder1 is an underannotator and Coder2 an overannotator. As clear from the plots,
  - When Tr is increasingly annotated by Coder2 (an overannotator), precision suffers somehow (along with more TPs there are also more FPs), but this is compensated by an increase in recall;
  - When Tr is increasingly annotated by Coder1 (an underannotator), recall drops substantially (due to fewer TPs), and this drop is not compensated by the stability of precision.
- An approximate randomization test (ART) confirms that the drop in F_1 is statistically significant in Batch1 but not in Batch2:

ART	LC-CRFs	
	F_1^μ	F_1^M
Batch1	0.0859	0.6207
Batch2	0.0001	0.0001

- The preliminary indications are thus that **low-quality training data are less of a problem if the training data annotator is an overannotator**

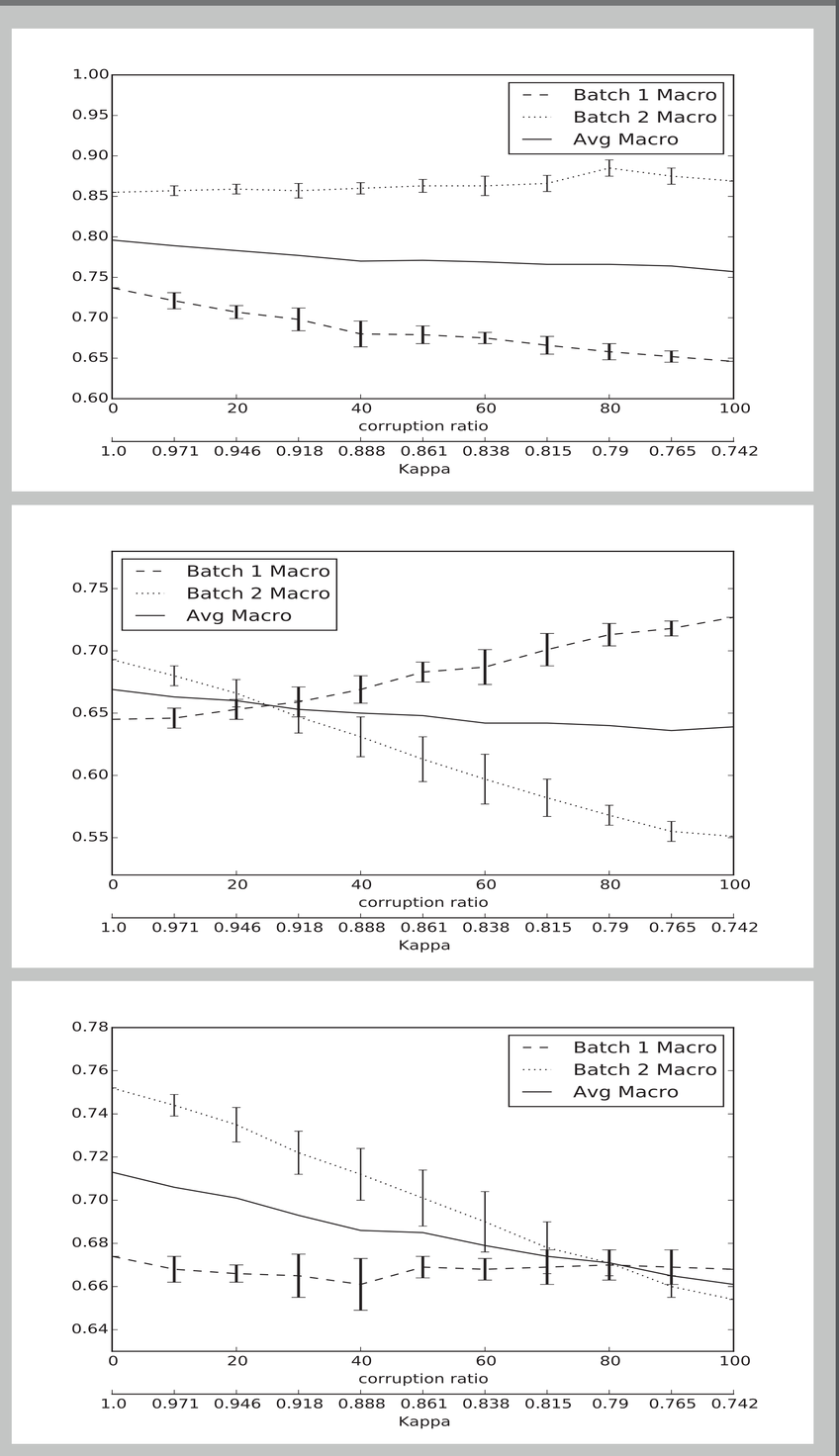

Figure: Macroaveraged precision (top), recall (mid), and F_1 (bottom) as a function of the corruption ratio λ for the LC-CRFs case.

Future work

- More experiments (and more datasets with double annotations) needed to confirm the above results.
- Future experiments also need to test situations characterized by lower levels of intercoder agreement (e.g., junior coders, crowdsourcers, etc.).

More details in ...

- D. Marcheggiani and F. Sebastiani. On the Effects of Low-Quality Training Data on Information Extraction from Clinical Reports. *arXiv:1502.05472 [cs.LG]*

http://www.qcr.qa/our-people/fabriziosebastiani fsebastiani@qf.org.qa