The Augmented Social Scientist

How to Automatically Annotate Millions of Texts with Human-Level Accuracy

Salomé Do[1,3], Étienne Ollion[2], Rubing Shen[1,2]

[1]Sciences-Po (Medialab), [2]CNRS (CREST), [3]ENS (LATTICE)

A Problem of Abundance

- **An Avalanche of Digital Data**
  - Born digital data
  - Digitized data
  - Including loads of textual data

A Problem of Abundance

- **How to extract meaning from this large trove?**
  - And old question, two responses

A Problem of Abundance

- The massive availability of textual data
- **How to extract meaning from this large trove?**
  And Old Question, two classic responses
  - Human annotation
    - Researcher, Research assistants, and now microworkers (Appen, AMT, TaskRabbit...)
    - Issues: costs, ethics, and even quality.

A Problem of Abundance

- The massive availability of textual data
- **How to extract meaning from this large trove?**
  And old question, two classic responses
  - Human annotation
  - Quantitative Text Analysis
    "Distant reading", from Bible indexes to Machine Learning
    - Merits and limits well-known
    - But could not until now reach the same level of precision as humans

A Problem of Abundance

- The massive availability of textual data
- What to do with Massive Textual Data?
- **Ideally, we could create an in-silico replica of an expert**
  - Training a model to replicate our own coding
  - This is what supervised methods are for
  - But until recently, relatively disappointing results

A Problem of Abundance

- The massive availability of textual data
- How to extract meaning from this large trove?
- Ideally, we could create an in-silico replica of an expert

- **This is exactly the promise of Large Language Models**
  - Recent developments (LLMs) claim to do away with this problem
  - But can they really do as well as humans?
  - Still begs questions: Amount of training data? Role of quality?

A Problem of Abundance

- **Two questions**
  - Can a social scientist train an efficient algorithm to replicate her fine-grained annotations (or is it too long/not as good)?
  - What is the trade-off between expert and profane annotation?
- **An Experiment**

The Question

- Assessing the rise of "strategic news coverage"
  - Political games over political measures
  - Revelation of backstage maneuvers
  - Evocation of the strategies of politicians

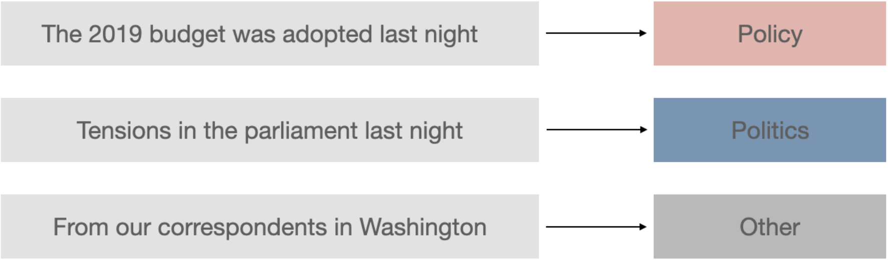
Data and Indicators

- All articles about politics from the French daily *Le Monde*
  - High-brow newspaper
  - Long Reluctant to adopt SNC (Kaciaf, 2013), yet did it (Saitta, 2005)

Years	Number of articles	Number of words	Number of journalists
1945-2018	61,511	38,497,810	113

Data and indicators

- Two tasks
  - **Task 1 - Policy *vs.* Politics**
    Content of a measure *vs.* Actions of politicians
    Complexity: high



Data and Indicators

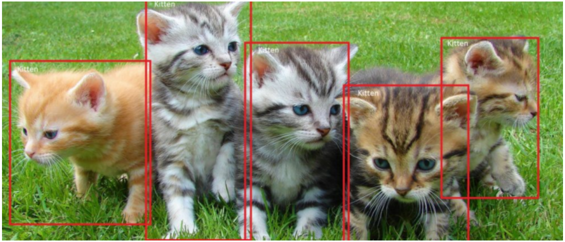
- Two tasks
  - **Task 2 - "Unattributed Quotes"**
    Prompts introducing unattributed speech
    Complexity: average+

A	source	close	to	power	confirmed	that	the	vaccine	won't	be	available	until	June
1	1	1	1	1	1	1	0	0	0	0	0	0	0

This	is	not	our	plan,	an	unnamed	official	told	government	reporters
0	0	0	0	0	1	1	1	1	1	1

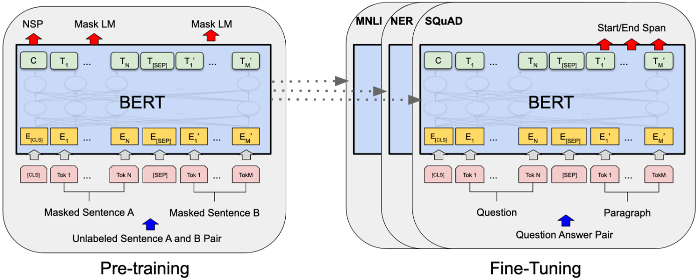
Intuition

- Supervised ML applied to text
  - Training an algorithm to mimic human annotation



Language Models

- Language models : Large pre-trained neural networks using Transformers such as BERT (Devlin et al., 2019) and CamemBERT for French (Martin et al., 2020)
- Pre-training : self-supervised "Masked Language Model" task, on French corpus OSCAR
- Fine-tuning : our tasks of text classification (Policy/Politics) and sequence labelling (Unattributed).



Design of the experiment

Annotators:

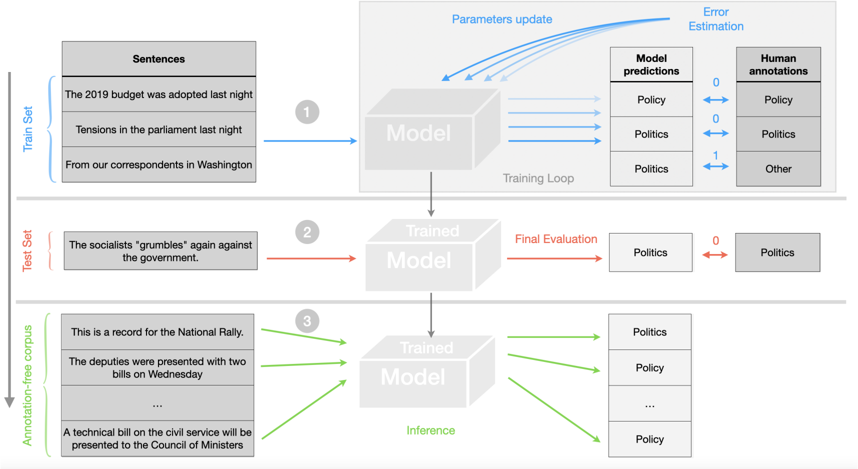
- **Social Scientist (SS)** An expert in her field. Often designs the indicators. Her time is limited.
  In this case: one of the authors of the paper.
- **Research assistants (RAs)** Trained, qualified students, but not experts. Interactions with the researcher.
  In this case: 3 Master's level RAs carefully trained by us.
- **Microworkers (MW)** Limited training, limited connections to the researcher.
  In this case: 34 BA students from a class. [Note: most likely better than gig workers]

Design of the Experiment

Experiment:

- For each task
  - A sample is annotated by a given group, for model training (*train set*, less than 1% of the whole corpus)
  - Experts defines a *gold standard* to validate the annotations (human or not)

Design of the Experiment



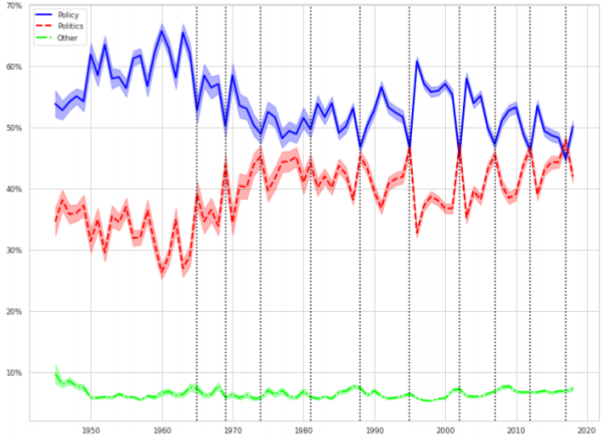
Design of the Experiment

		Social Scientist	**Research Assistants**	**Microworkers**
Number of annotators		1	3	34
Level of expertise		High	Moderate	Low
Train set size	Task 1	63 articles 383 000 characters (0.12% of the corpus)		
	Task 2	6274 excerpts (3 sentences of an article) 3,1 millions characters (0.91% of the corpus)		
Time spent (in minutes)	Task 1	480	1051	1243
	Task 2	2220	1869	2654

Result 1: Training

A qualitative assessment of the predictions

Politics vs. policy in *Le Monde*



Result 1: Training

	Policy vs. Politics	Unattributed
Human - Microworkers	0.65	0.7
Human - RAs	**0.80**	**0.86**

Table: F-1 Score for human annotation

Comparison to a Gold Standard annotated with care by experts

Result 1: Training

	Policy vs. Politics	Unattributed
Human - Microworkers	0.65	0.70
Human - RAs	**0.80**	**0.86**
"Classic" supervised models	0.67	0.41
(LSTM, SVM)	[0.671, 0.673]	[0.390, 0.437]
Augmented Social Scientist	0.78	0.82
(camemBERT)	[0.781, 0.792]	[0.816, 0.834]

Table: F-1 Score for human annotation vs. Model trained by the expert

Result 1: Training

A qualitative assessment of the predictions

Manual classification of the Unattributed predictions

Type	**Frequency**
(Quasi-) agreement	76%
Partial agreement	2%
In gold standard, but not predicted (= false negative)	10%
Predicted correctly by the algorithm, but not noticed by the expert	8%
Predicted incorrectly (= false positive)	4%

Result2: Trade-Off

What is the role of expertise in training a model?

Result2: Trade-Off

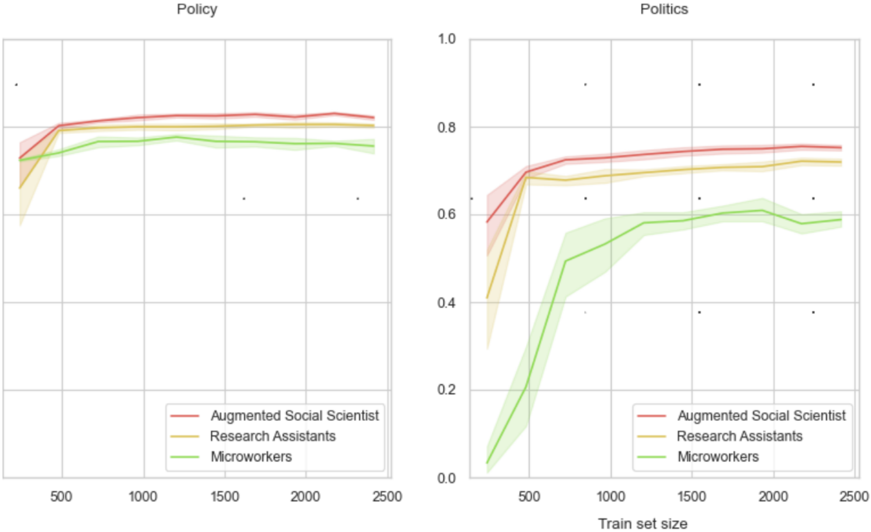


Figure: Sample-efficiency curves (F-1 score)

Result2: Trade-Off

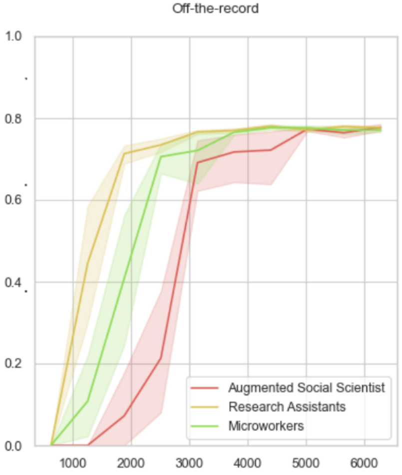


Figure: Sample-efficiency curves (F-1 score)

An Immense Promise

- More than satisfactory results, in a limited amount of time
  Time spent (in minutes):

	Task 1	Task 2
Social Scientist	480	2220
Research Assistants	1218 mean=406 sd=191	2190 mean=730 sd=407
Microwokers	2014 mean=59 sd=24	2851 mean=84 sd=36

Keep in mind: There are even more ways to cut down on annotation

An Immense Promise

- An Improvement with respect to classic methods
  - Qualitative Methods
  - Outsourced hand annotation
  - Non-supervised models
- Ability to fully annotate a vast data set
  - Comprehensive AND fine-grained annotation (at the level of the article, or even below)
  - Forces conceptual clarification
  - Avoid classic pitfalls of hand-annotation (fatigue effect, learning effect) (Rousson et al., 2002)
  - Several good validation criteria

Limitations and Challenges

- Computer time and hardware
  - Hard without a GPU
  - Colab and its problems
  - Ask your institution for resources
- When to use it? And who should annotate?
- Still cannot replace humans, in many ways.

Conclusion

SML, STL, and Human Augmentation

- And old debate: machines to replace, or machine to augment individuals?
- Doug Engelbart, the Internet and "The Augmentation Research Lab" "Increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems." (Engelbart, *Augmenting Human Intellect*, 1962).

References I

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. Summary Report AFOSR-3223 under Contract AF 49(638)-1024, SRI Project 3578 for Air Force Office of Scientific Research. Menlo Park, Ca., Stanford Research Institute.

Kaciaf, N. (2013). *Les pages "Politique". Histoire du journalisme politique dans la presse française (1945-2006)*. Presses universitaires de Rennes.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. Association for Computational Linguistics.

Rousson, V., Gasser, T., and Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in medicine*, 21(22):3431–3446.

Saitta, E. (2005). *Le Monde*, vingt ans après. *Réseaux*, 131(3):189–225.