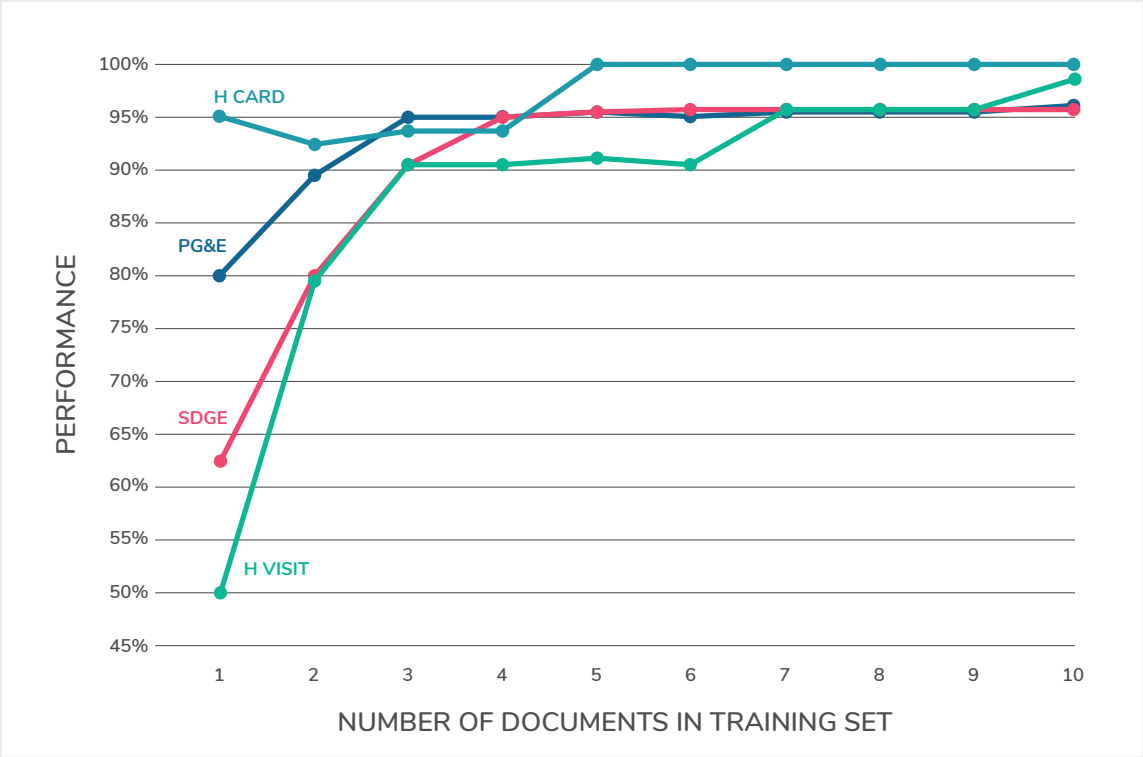
GLYNT.AI

LOW SHOT LEARNING IN ACTION

Unstructured Data Extraction Results from Four Document Sets



Results from GLYNT data extraction.
Training on 1 – 10 documents. Extraction on 40 documents.

Introduction

Low Shot Learning produces amazing results. As the graph above shows, Low Show Learning added to the GLYNT system achieves excellent performance with just a handful of documents in the training set. This technical note details how the results were obtained and provides insights on how Low Shot Learning works in practice.

What is particularly important is the high performance level, above 95%, achieved by the GLYNT system with Low Shot Learning. As Andrew Ng and other leading AI practitioners have noted, businesses are slow to adopt AI solutions when performance is in the mid-80s. There are too many false positives and errors. But when performance levels are greater than 95%, corporate adoptions move forward.

GLYNT with Low Shot Learning needs little staff time to achieve the required performance level. This is a significant change in the AI landscape. No specialized software engineers are needed, as office workers with modest tech skills can easily accomplish the tasks that produced the results in this paper. Low Shot Learning reduces time and effort, compute costs, staff expense and project risk. Instead of spending months of time to get an uncertain result, GLYNT with Low Shot Learning produces excellent results in under an hour.

[1]David Kiron, keynote address at AI World, December 2017

The Data

Documents from four sources were used in this study and 50 documents were obtained from each source. 10 documents were made available to GLYNT's training feature, and 40 documents were kept for the holdout data set. 8 - 11 fields were extracted per document set. The four document sets are:

- Healthcare insurance cards
- Visit summaries from a healthcare clinic
- Utility bills from Pacific Gas & Electric
- Utility bills from San Diego Gas & Electric

With strict privacy laws in healthcare (known as HIPPA), synthetic data sets were developed for the insurance cards and clinic visit summaries. An image of each was taken from the internet, and then the data elements were randomly varied across 50 synthetic document images. It is expected that these document sets will have fewer OCR errors, but no particular advantage in the extraction of specific data fields through GLYNT's machine learning system. In practice, the results of the two synthetic document sets are comparable to results from non-synthetic document sets.

The data items highlighted in blue are the fields to be extracted from this document. While the healthcare insurance card is a very compact document, the utility bills are 3 - 5 pages each and data items from several pages per bill were used in this study.

Sample United Healthcare Insurance card

The Method

The following method was used for each document set:

- Create a database from the 50 documents of the data values for the desired fields, e.g. Ground Truth
- Prepare an inventory of fields by document
- Separate the documents into a Training Set (10 documents) and a Holdout Set (40 documents). The Training Set is curated so that all of the desired fields appear in the group of documents
- Train 10 machine learning models for each document set. The first model is based on a Training Set of one document. The second model is basted on a Training Set of two documents, and so on
- Use each model to extract data from the Holdout Set, producing extraction results 1 through 10
- For each set of extraction results, report the Precision, Recall and F1 scores

Summary of Model Performance

Data extraction results for each iteration of the machine learning model are summarized in the following report. The example below is for the Healthcare Visit Summary, with four documents in the Training Set and 40 documents in the Holdout Set.

		PERFORMANCE			SCORES		
Field Name	**Field Type**	**Correct**	**Incorrect**	**Missing**	**Precision**	**Recall**	**F1**
Appointment Date	Date	33	0	7	100%	83%	90%
BMI	Numeric	39	1	0	98%	100%	99%
Clinic Name	Phrase	32	8	0	80%	100%	89%
Diagnoses1	Phrase	29	1	10	97%	74%	84%
Diagnoses2	Phrase	30	1	9	97%	77%	86%
Provider	Alphanumeric Code	33	0	7	100%	83%	90%
Reason for Visit	Phrase	33	7	0	83%	100%	90%
Weight	Numeric	38	0	2	100%	95%	97%
Total		**267**	**18**	**35**	**94%**	**88%**	**91%**

Note that a variety of field types are extracted. For example, one of the Diagnoses text phrases is ABDOMINAL DISCOMFORT, LEFT QUADRANT [894.01], and a sample Provider code is T-15. Thus Low Shot Learning is being tested against simple date fields as well as complex phrases.

The middle three columns of the table show the performance counts by field for the 40 documents in the Holdout Set. There are three possible outcomes:

Correct: If the data value in the Ground Truth data set exactly matches the data value extracted by the trained model, the result is labeled Correct. Exact match requires that the two data items have precisely the same characters, eg a 100% character-level match.

Incorrect: If the data value extracted by the trained model does not exactly match the the data value in the Ground Truth data set, then the result is labeled Incorrect.

Missing: If the GLYNT machine learning system is not highly confident it is producing the correct data item for the field, it is set up to suppress the return of the data item. This leads to an empty field, which is labeled as Missing.

Note that the counts in the three performance categories sum to the total number of documents in the Hold Out set. For some document sets, there are a variety of fields, and not all fields are shown on all documents. In this event, additional information is included in the Report to provide a complete inventory of the field-level data shown on the documents and the formulae for the scores are adjusted appropriately. For simplicity, the Healthcare Visit Summary data set is shown here, in which every document has every field present.

Turning to the final three columns the three scores are defined as follows:

Precision: Correct / Total Documents
Recall: (Total Documents - Missing)/Total Documents
F1: [2 x Precision x Recall] / [Precision + Recall]

While Precision focuses on the common concern, "Is this result accurate?", Recall answers the question, "What percent of the desired data fields are provided by this training model?". The F1 score is a blend of the two measures.

Comparison of Training Results by Document Set

The following table shows the performance scores for all models and all data sets. There are several items to note.

First, Precision is quite high throughout, even with just one document in the training set. Recall improves dramatically with additional training, but Precision is high the start; there is little upward movement with additional documents in the training set.

Second, there is a bit of noise in the results. For example, the Precision results for Visit Summaries starts at 100%, then decreases to 94% when more training documents are added, and then climbs back to 99%. The Holdout Set is the same for each of these results, so the the variation in results must arise from the changes in the Training Set. The issue is that Low Shot Learning is a very tight model, and small variations in Training Data lead to small variations in results. However, these disappear as the Training Set increases beyond just 2 – 5 documents.

	PG&E			SDG&E			Insurance Cards			Visit Summaries			Average		
# of documents in training sets	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
1	100	66	**80**	96	47	**63**	100	90	**95**	100	33	**50**	99	59	**72**
2	99	81	**89**	96	69	**80**	99	88	**93**	99	66	**79**	98	76	**85**
3	99	91	**95**	97	86	**91**	100	88	**94**	96	87	**91**	98	88	**93**
4	99	92	**95**	97	93	**95**	100	88	**94**	94	88	**91**	98	90	**94**
5	98	94	**96**	97	93	**95**	100	100	**100**	94	90	**92**	97	94	**96**
6	97	93	**95**	97	95	**96**	100	100	**100**	96	86	**91**	98	94	**95**
7	98	94	**96**	97	95	**96**	100	100	**100**	97	96	**96**	98	96	**97**
8	98	95	**96**	97	95	**96**	100	100	**100**	97	96	**96**	98	97	**97**
9	98	96	**96**	97	95	**96**	100	100	**100**	97	96	**96**	98	97	**97**
10	98	97	**97**	97	95	**96**	100	100	**100**	99	98	**98**	99	98	**98**

Results are in Percent

Third, as the Average results show very clearly, adding training documents improves Recall in the GLYNT system. Not much tradeoff between Precision and Recall is possible, thus, parameter tuning in attempt to balance Precision and Recall for maximum F1 score is unnecessary. This was an unexpected result given that parameter tuning is a product feature built into GLYNT. However, it has not been needed since the implementation of Low Shot Learning.

Fourth, the synthetic document sets and the utility bill document sets achieved remarkably similar performance results. Real-world documents have more varied layouts, poor quality document scans, and so on. There are a number of reasons to expect clean synthetic documents to perform better than their messier real world counterparts. Yet, the results indicate that with a just handful of documents, training is excellent. At 10 documents in the training set, the real world documents are at or above 95% in both Precision and Recall, and with F1 scores only 2 - 3 points below the F1 scores of the synthetic documents.

Fifth, GLYNT with Low Shot Learning is incredibly powerful and dominates other machine learning solutions. Compare GLYNT's performance and required training data sets with these data from other machine learning solutions for document classification and data extraction:

Dropbox: 87% accuracy after training on more than 300,000 invoices
Cloudscan: 84% accuracy after training on more than 326,000 invoices
Jatana.ai: 85% accuracy after training on more than 14,000 documents
GLYNT: 98% accuracy after training on less than 10 documents

Finally, observe that the graph shown at the start of this article shows the results for the F1 score. Typically machine learning results are displayed in terms of Precision only. But the precision results for GLYNT with Low Shot Learning start high and stay that way. Without context they are confusing. So, we are presenting the F1 score in graphs. GLYNT's overall performance matches or exceeds that of other machine learning and OCR systems, yet during the coming months we'll be focusing on improving Recall. Our goal is make GLYNT's F1 score results also difficult to interpret!

Conclusion

Every industry has treasure troves of data trapped in pdfs, scans and faxes. Liberating this data improves analytics and increases operational efficiency. GLYNT with Low Shot Learning makes the extraction of unstructured data faster, better and cheaper than ever before.

AI is a fast-moving industry, and GLYNT has a key role to play, the liberation of unstructured data. Imagine a world in which the data-hungry applications – such as AI for security, marketing or operations – also used Low Shot Learning. "Low Shot Squared" is quite the accelerant, dramatically changing how quickly AI tools could be tested and deployed. Liberating trapped data with GLYNT is the first step.