Finetuning

Jordan Boyd-Graber

University of Maryland

DSPy Example

Plan for Today

- Alternative to Prompt Engineering
- Example of working with DSPy program
  - RAG
  - String and Float Output
  - Custom Objective Function

RAG

LLM generates a query to find text that will help guesser create a better guess.

Guess

LLM looks at context to get best possible answer.

Calibration

LLM looks at guess and says how confident the answer is.

Setup Retriever

```
def init_retriever(self, name,
                   model_filename, topk=1):
    logging.info(("Loading retriever %s as %s "
                  "(top k=%i)") %
                   (model_filename, name, topk))
    from tfidf_guesser import TfidfGuesser
    retriever = TfidfGuesser(model_filename)
    self.retrievers[name] = retriever
    self.retrievers[name].load()
    if self.topk is None:
        self._topk = topk
    else:
        assert self._topk == topk
```

Generate a Query

```
class QueryGenerator(dspy.Signature):
    question: str = dspy.InputField()
    query: str = dspy.OutputField()
```

Example

Input: `He wrote about being indigent in two European capitals in "Down and Out in London and Paris”.`
Output: `Author Down and Out in London and Paris`

Generate a Guess

```
self.guess_generator = dspy.ChainOfThought("question,conte
```

Example

Input:

```
{
  question: ' He wrote about being indigent in two Europ
  query: 'Author Down and Out in London and Paris',
  context: {'Jack London: The house of his birth burned
}
```

Output: {guess: "Samuel Clemens”}

Generate a Confidence

```
class ConfidenceGenerator(dspy.Signature):
    question: str = dspy.InputField()
    query: str = dspy.InputField()
    context: str = dspy.InputField()
    guess: str = dspy.InputField()
    confidence: float = dspy.OutputField()
```

Example

Input:

```
{
  question: ' He wrote about being indigent in two Europ
  query: 'Author Down and Out in London and Paris',
  context: {'Jack London: The house of his birth burned
  guess: 'Samuel Clemens'
}
```

Output: {confidence: 0.27}

How good is our guess?

```python
def validate_answer(example, pred, trace=None):
    from eval import rough_compare

    correct = rough_compare(example.answer, pred.guess)

    if correct:
        return 1 + pred.confidence ** 2
    else:
        return - (pred.confidence ** 2)
```

How good is our guess?

```python
def validate_answer(example, pred, trace=None):
    from eval import rough_compare

    correct = rough_compare(example.answer, pred.guess)

    if correct:
        return 1 + pred.confidence ** 2
    else:
        return - (pred.confidence ** 2)
```

Not Perfect

- Improve RAG query: Answer in the context
- Improve Guess: Context came from question with correct answer
- Improve Calibration: Reasoning for guess, consistent with context

Full Program

```
class FullResult(dspy.Signature):
  guess: str = dspy.OutputField()
  confidence: float = dspy.OutputField()
  context: str = dspy.OutputField()
```

- Repeating previous components
- Want to keep it around (i.e., for metric, downstream buzzer)

Full Program

```
def forward(self, question, **kwargs):
   query = self.query_generator(question=question).query
   context = ""
   for retriever in self.retrievers:
       context += "%s: %s" % \
          (retriever, str(self.retrievers[retriever](query, self._topk)))

   guess = self.guess_generator(question=question, context=context)
   confidence = self.confidence_generator(question=question, query=query,
                                          context=context, guess=guess)
   return FullResult(guess=guess.answer, context=context,
                     confidence=confidence.confidence)
```

Create Example Object

```
@staticmethod
def create_dataset(questions, answers):
    return [dspy.Example(question=x, answer=y) for
            x, y in zip(questions, answers)]
```

Evaluate Program

```
od = OllamaDspy(filename="models/ollama_guesser")
with gzip.open("data/qanta.guessdev.json.gz") as infile:
  questions = json.load(infile)
  q_field = [sent_tokenize(x["text"])[:-1]
             for x in questions[:129]]
  a_field = [x["page"] for x in questions[:129]]
  dev = od.create_dataset(q_field, a_field)
  evaluator = Evaluate(devset=dev, num_threads=5,
                       display_progress=True,
                       display_table=10)
```

So many magic numbers!

- -1: Including the last sentence is too easy (could do runs)
- 129: Our standard dev set has 1129 examples, use 1000 for teleprompter
- 5: What my computer could handle (more caused too many files open error)

No optimization: -2.15

Question	Answer	Guess	Score
Robert Walker argued that failing to take th...	Texas annexation	Anson Jones proposed...	-0.578
In one of this director's films, the opening...	Martin Scorsese	Travis Bickle	-0.902
Along with orbitons and holons, quasiparticl...	Spin	Spin	1.911
This singer instructs "put me onto your blac...	Lana Del Rey	Lana Del Rey	1.902
By processing one etching through four stage...	Rembrandt	Johannes Vermeer	-0.902
Cristobal de Morales composed a work in this...	Mass (music)	I cannot answer...	-0.003
Metabolism of this molecule is disturbed by ...	DNA	DNA	1.902
This ruler's dream of his son's death by an ...	Croesus	Cyrus the Great	-0.916
This author describes the title event of one...	Gerard Manley Hopkins	Rainer Maria Rilke	-0.912
In this television show, a character smells ...	Hannibal	Hannibal	1.574

Setup Teleprompter

```
from dspy.teleprompt import MIPROv2
self._teleprompter = MIPROv2(
        metric=validate_answer,
        num_threads=4,
        auto='medium',
)

self._optimized = self._teleprompter.compile(
    CalibratedRAG(),
    trainset=dataset,
)
```

Query Prompts

- Given the fields ‘question’, produce the fields ‘query’.
- Generate a query based on the provided question. The query should be suitable for use in a retrieval system to find relevant context for answering the question.
- Transform the given question into a retrieval query suitable for the RAG pipeline. Focus on extracting the key terms and concepts from the question to effectively guide the retrieval process.
- Generate a query based on the input question.
- Imagine you are a highly-ranked intelligence analyst tasked with rapidly identifying the origins of a coded message intercepted from a suspected terrorist cell. . . .
- You are a Retrieval-Augmented Generation (RAG) system expert tasked with generating the initial retrieval query for a RAG pipeline. Given the input question, produce a concise and effective query . . .

Guess Prompts

- Given the fields ‘question‘, ‘context‘, produce the fields ‘answer‘.
- You are a philosophical question-answering system designed to tackle complex questions related to consciousness and the mind-body problem. . .
- You are a philosophical assistant specializing in consciousness studies. Given the user’s question and the provided context, formulate a concise answer representing your understanding of the topic, mir found in philosophical discourse. Output the answer.
- You are a sophisticated AI assistant designed to provide answers based on a given question and associated context. Your task is to analyze the provided question and context, and then generate . . .
- Based on the input question and provided context, generate a concise answer. Utilize a step-by-step reasoning process if helpful, and output the final answer.
- You are a question answering assistant using a Retrieval-Augmented Generation (RAG) system. You will be given a ‘question‘ and a ‘context‘. Your task is to generate a concise and accurate ‘answer‘ . . .

Confidence Prompts

- Given the fields ‘question‘, ‘query‘, ‘context‘, ‘guess‘, produce the fields ‘confidence‘.
- Imagine you are a leading neurophilosopher tasked with advising a newly established space program. The program aims to develop artificial consciousness in extraterrestrial robots. Your task is to . . .
- . . . Accuracy is paramount a misjudgment could jeopardize the entire project!
- Given the fields ‘question‘, ‘query‘, ‘context‘, ‘guess‘, produce the fields ‘confidence‘. Specifically, formulate a confidence score (a float between 0.0 and 1.0) representing the models certainty in. . .
- You are a highly skilled philosophical analyst tasked with evaluating the reliability of an answer generated by a Retrieval-Augmented Generation (RAG) system. The system has provided an initial . . .
- For a prestigious international debate competition, you must accurately identify the key figure responsible for a pivotal historical eventspecifically, the signing of the Treaty of Versailles. You . . .

Final Prompts

Query

Generate a query based on the given question.

Final Prompts

Query

Generate a query based on the given question.

Guess

You will be provided with a question and a context. Your task is to generate a concise and accurate answer based on the provided context. Focus on extracting the most relevant information and presenting it clearly. Output the answer field directly.

Final Prompts

Query

Generate a query based on the given question.

Guess

You will be provided with a question and a context. Your task is to generate a concise and accurate answer based on the provided context. Focus on extracting the most relevant information and presenting it clearly. Output the answer field directly.

Confidence

Imagine a catastrophic event – a sudden, unexplained disappearance of all historical records – has occurred. You are tasked with using the provided question, query, context (representing a fragmented attempt to reconstruct the past), and a generated “guess” to estimate the confidence of your response. This is a high-stakes scenario demanding precise reasoning and a robust assessment of uncertainty . . .

Post-Optimization: 5.43

Question	Answer	Guess	Score
Robert Walker argued that failing to take th...	Texas annexation	Texas Annexation	1.902
In one of this director's films, the opening...	Martin Scorsese	Martin Scorsese	1.980
Along with orbitons and holons, quasiparticl...	Spin	Spin	1.960
This singer instructs "put me onto your blac...	Lana Del Rey	Sia	-0.960
By processing one etching through four stage...	Rembrandt	James McNeill Whistler	-0.960
Cristobal de Morales composed a work in this...	Mass	Mass	1.902
Metabolism of this molecule is disturbed by ...	DNA	DNA	1.980
This ruler's dream of his son's death by an ...	Croesus	Croesus	1.960
This author describes the title event of one...	Gerard Manley Hopkins	John Keats	-0.960
In this television show, a character smells ...	Hannibal	Westworld	-0.960

Post-Optimization: 5.43

Question	Answer	Guess	Score
Robert Walker argued that failing to take th...	Texas annexation	Texas Annexation	1.902
In one of this director's films, the opening...	Martin Scorsese	Martin Scorsese	1.980
Along with orbitons and holons, quasiparticl...	Spin	Spin	1.960
This singer instructs "put me onto your blac...	Lana Del Rey	Sia	-0.960
By processing one etching through four stage...	Rembrandt	James McNeill Whistler	-0.960
Cristobal de Morales composed a work in this...	Mass	Mass	1.902
Metabolism of this molecule is disturbed by ...	DNA	DNA	1.980
This ruler's dream of his son's death by an ...	Croesus	Croesus	1.960
This author describes the title event of one...	Gerard Manley Hopkins	John Keats	-0.960
In this television show, a character smells ...	Hannibal	Westworld	-0.960

Post-Optimization: 5.43

Question	Answer	Guess	Score
Robert Walker argued that failing to take th…	Texas annexation	Texas Annexation	1.902
In one of this director's films, the opening…	Martin Scorsese	Martin Scorsese	1.980
Along with orbitons and holons, quasiparticl…	Spin	Spin	1.960
This singer instructs "put me onto your blac…	Lana Del Rey	Sia	-0.960
By processing one etching through four stage…	Rembrandt	James McNeill Whistler	-0.960
Cristobal de Morales composed a work in this…	Mass	Mass	1.902
Metabolism of this molecule is disturbed by …	DNA	DNA	1.980
This ruler's dream of his son's death by an …	Croesus	Croesus	1.960
This author describes the title event of one…	Gerard Manley Hopkins	John Keats	-0.960
In this television show, a character smells …	Hannibal	Westworld	-0.960

Post-Optimization: 5.43

Question	Answer	Guess	Score
Robert Walker argued that failing to take th. . .	Texas annexation	Texas Annexation	1.902
In one of this director's films, the opening. . .	Martin Scorsese	Martin Scorsese	1.980
Along with orbitons and holons, quasiparticl. . .	Spin	Spin	1.960
This singer instructs "put me onto your blac. . .	Lana Del Rey	Sia	-0.960
By processing one etching through four stage. . .	Rembrandt	James McNeill Whistler	-0.960
Cristobal de Morales composed a work in this. . .	Mass	Mass	1.902
Metabolism of this molecule is disturbed by . . .	DNA	DNA	1.980
This ruler's dream of his son's death by an . . .	Croesus	Croesus	1.960
This author describes the title event of one. . .	Gerard Manley Hopkins	John Keats	-0.960
In this television show, a character smells . . .	Hannibal	Westworld	-0.960

Feature Engineering is New Again

- You should not do prompt tuning by hand
- Figure out what you care about and measure it
- Understand your data and problem
- Next steps

Feature Engineering is New Again

- You should not do prompt tuning by hand
- Figure out what you care about and measure it
- Understand your data and problem
- Next steps
    - Get more inputs for confidence estimation: multiple muppet models, reasoning chains
    - Derive more features from context
    - Tune RAG system: bigrams, higher recall, add Wikipedia