

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

Constructing a Knowledge Base on Aging

An Automated Approach

Mark Farrell

Bioinformatics Researcher

Center for Research and Education on Aging
Lawrence Berkeley National Laboratory
University of California, Berkeley

September 4th, 2014

Outline

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

1 About Me

2 Automatically Constructing Knowledge Bases

3 Extracting Facts in a Structured Format

4 Results & Discussion

Who I am

Mark Farrell

Constructing a
Knowledge
Base on Aging

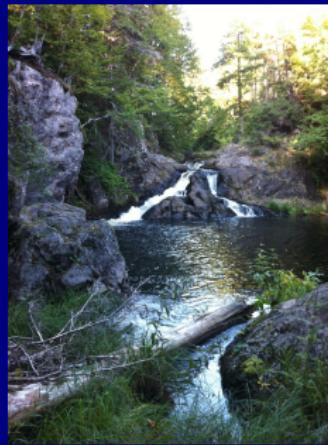
Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion



Where I'm from

Atlantic Canada

Constructing a
Knowledge
Base on Aging

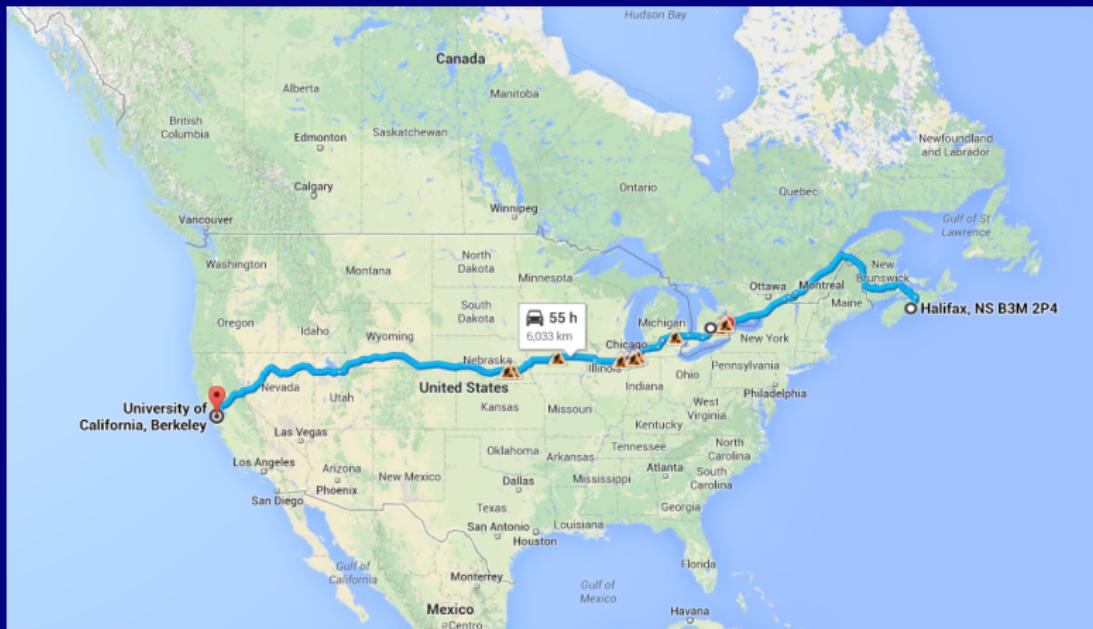
Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion



What I'm Studying

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

Year 2

Program Bachelor of Computer Science

Faculty Mathematics

Institution University of Waterloo

Location Waterloo, Ontario, Canada



How I Became a Researcher at CREA

Prof. Garan Emailed the University of Waterloo Computer Science Club

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion



Why was I Interested in Research at CREA?

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

Hobby	CREA Research
Reverse Engineering	Natural Language Processing
Game Development	Artificial Intelligence
Health	Bioinformatics

Outline

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

1 About Me

2 Automatically Constructing Knowledge Bases

3 Extracting Facts in a Structured Format

4 Results & Discussion

Overview

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- CREA is constructing a knowledge base to study and understand the human aging process.

Overview

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- CREA is constructing a knowledge base to study and understand the human aging process.
- New discoveries are published quickly and in large volume.

Overview

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- CREA is constructing a knowledge base to study and understand the human aging process.
- New discoveries are published quickly and in large volume.
- It is infeasible to construct the knowledge base by hand.

Overview

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- CREA is constructing a knowledge base to study and understand the human aging process.
- New discoveries are published quickly and in large volume.
- It is infeasible to construct the knowledge base by hand.
- Working on software to construct the knowledge base automatically.

Introduction

How to Automatically Construct the Knowledge Base

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Routinely search for keywords related to aging, dowloading text articles from sources like PubMed.

Introduction

How to Automatically Construct the Knowledge Base

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Routinely search for keywords related to aging, dowloading text articles from sources like PubMed.
- Build a spam filter to get rid of non-scientific sentences.

Introduction

How to Automatically Construct the Knowledge Base

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Routinely search for keywords related to aging, dowloading text articles from sources like PubMed.
- Build a spam filter to get rid of non-scientific sentences.
- Extract scientific facts from the sentences and save them in a structured format.

Introduction

How to Automatically Construct the Knowledge Base

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Routinely search for keywords related to aging, dowloading text articles from sources like PubMed.
- Build a spam filter to get rid of non-scientific sentences.
- Extract scientific facts from the sentences and save them in a structured format.
- Provide a graphical interface that allows users to search and otherwise explore the knowledge base.

Summary of Progress

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Devised and implemented the method for finding simple facts in sentences, extracting them in a structured format.

Summary of Progress

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Devised and implemented the method for finding simple facts in sentences, extracting them in a structured format.
- Began work on a web viewer for the knowledge base.

Outline

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

1 About Me

2 Automatically Constructing Knowledge Bases

3 Extracting Facts in a Structured Format

4 Results & Discussion

Tokenization

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Input a text document and read it, one sentence at a time.

Example: Tokenization

```
scala> tokens("The man walks. The dog eats.")  
res0: List[String] = List(The man walks., The dog eats.)
```

Parsing

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- For each sentence, generate a constituent tree that describes its phrase structure.

Example: Parsing

```
scala> parse("The man walks the dog." )  
res0: Tree[String] = (ROOT  
  (S  
    (@S  
      (NP (DT The) (NN man))  
      (VP (VBZ walks)  
        (NP (DT the) (NN dog))))  
    (. .)))
```

Parsing Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

■ The University of Pennsylvania Treebank Project:

Parsing Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- The University of Pennsylvania Treebank Project:
 - Defines notation for constituent trees.

Parsing Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- The University of Pennsylvania Treebank Project:
 - Defines notation for constituent trees.
 - Parses sentences from the Wall Street Journal by hand.

Parsing Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- The University of Pennsylvania Treebank Project:
 - Defines notation for constituent trees.
 - Parses sentences from the Wall Street Journal by hand.
- The Berkeley Parser is software that guesses how to parse a sentence from the notation and examples specified by the Penn Treebank.

Compilation

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Extract facts from each constituent tree.

Example: Compilation

```
scala> compile("The man walks the dog.").shows  
res0: String = [<compound:walk(<atom:man>, <atom:dog>)>]
```

Compilation Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

Pattern match on the constituent trees. Define patterns for:

- 1 Extracting nouns from noun phrases (NP).

Compilation Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

Pattern match on the constituent trees. Define patterns for:

- 1 Extracting nouns from noun phrases (NP).
- 2 Extracting predicates and nouns from verb phrases (VP).

Compilation Method

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

Pattern match on the constituent trees. Define patterns for:

- 1 Extracting nouns from noun phrases (NP).
- 2 Extracting predicates and nouns from verb phrases (VP).
- 3 Extracting facts from complete clauses (S), making logical assertions with nouns and predicates.

Outline

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

1 About Me

2 Automatically Constructing Knowledge Bases

3 Extracting Facts in a Structured Format

4 Results & Discussion

Software Demonstration

A preview of CREA's knowledge base, compiled from PubMed abstracts.

Constructing a
Knowledge
Base on Aging

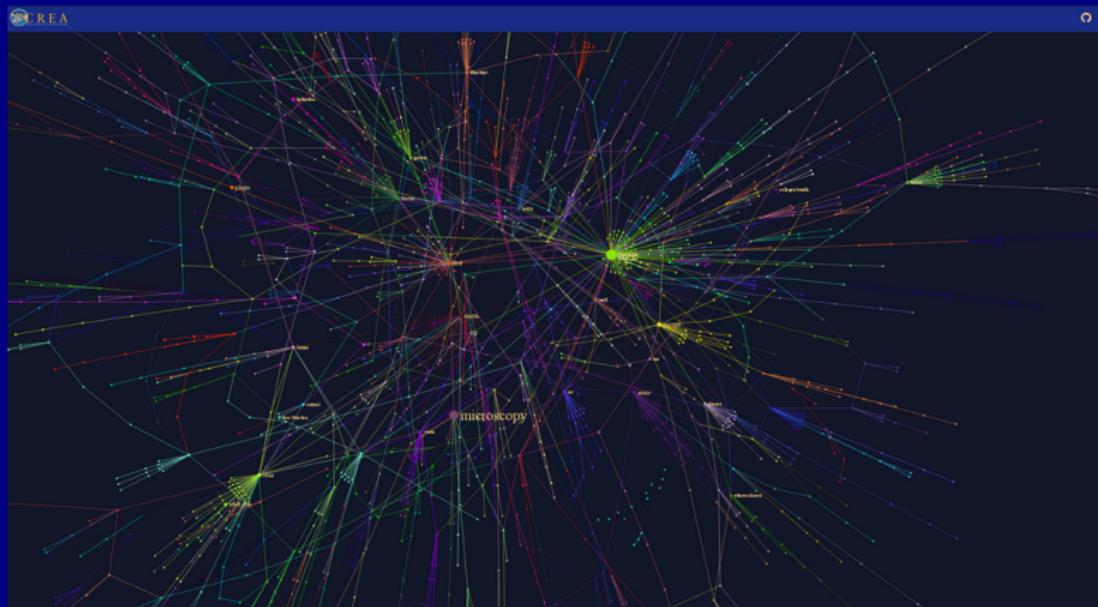
Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion



Performance

Parallelization

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- It is possible to extract facts from many sentences at the same time.

Accuracy

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Filter spam sentences from documents.

Accuracy

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Filter spam sentences from documents.
- The accuracy of the parser could be optimized:

Accuracy

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Filter spam sentences from documents.
- The accuracy of the parser could be optimized:
 - Should be trained to identify more nouns from the biomedical domain.

Accuracy

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Filter spam sentences from documents.
- The accuracy of the parser could be optimized:
 - Should be trained to identify more nouns from the biomedical domain.
- Define more patterns for extracting facts:

Accuracy

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Filter spam sentences from documents.
- The accuracy of the parser could be optimized:
 - Should be trained to identify more nouns from the biomedical domain.
- Define more patterns for extracting facts:
 - The software succeeds around 50% of the time.

Missing Features

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Support negated clauses and conditional logic.

Missing Features

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Support negated clauses and conditional logic.
- Facts can contradict each other:

Missing Features

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Support negated clauses and conditional logic.
- Facts can contradict each other:
 - Store the probability that is true as the weight of its edge on the knowledge base's graph.

Missing Features

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Support negated clauses and conditional logic.
- Facts can contradict each other:
 - Store the probability that is true as the weight of its edge on the knowledge base's graph.
- Scale and launch the software service.

Conclusion

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Demonstrated a method for automatically constructing CREA's knowledge base on aging.

Conclusion

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Demonstrated a method for automatically constructing CREA's knowledge base on aging.
- Showed how to extract facts from English text in the knowledge base's structured format.

Conclusion

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Demonstrated a method for automatically constructing CREA's knowledge base on aging.
- Showed how to extract facts from English text in the knowledge base's structured format.
- Discussed the need to improve software accuracy by lensing in on the biomedical domain.

Conclusion

Constructing a
Knowledge
Base on Aging

Mark Farrell

About Me

Automatically
Constructing
Knowledge
Bases

Extracting
Facts in a
Structured
Format

Results &
Discussion

- Demonstrated a method for automatically constructing CREA's knowledge base on aging.
- Showed how to extract facts from English text in the knowledge base's structured format.
- Discussed the need to improve software accuracy by lensing in on the biomedical domain.
- Suggested how the software implementation can be scaled for production usage.