Converting Fieldbooks to Databases

Talk given by Carsten Ehrler for the Project Seminar
"Text Mining for Historical Documents",
Computational Linguistics Department
Saarland University - 23.02.2009

Based on the publication: Sander Canisius and Caroline Sporleder. *Bootstrapping information extraction from field books*.
In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, pp. 827-836.

Introduction

"Sander Canisius and Caroline Sporleder. Bootstrapping information extraction from field books. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, pp. 827-836."

Introduction

Author: Canasius, Sander; Sporleder, Caroline

Title: Bootstrapping information extraction from field books

Type: Proceedings

Conference: Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)

Year: 2007

Location: Prague, Czech Republic

Page: 827-836

Overview

- Semi-structured documents
- Field-segmentation
- Field-segmentation methods
- Practical examples

Data Sources

Examples for semi-structured documents:

- apartment advertisements
- logs (e.g. archeological findings)
- business cards
- web-pages
- …

ten! 4 ZKDB, G-WC, Balkon, Miete 655 EUR zzgl. NK Tel.:0173/5303148

MG Viktorieastr. 2ZKDB Blk. 55m², ren. 280 + NK + Kaut. 0176/51150216

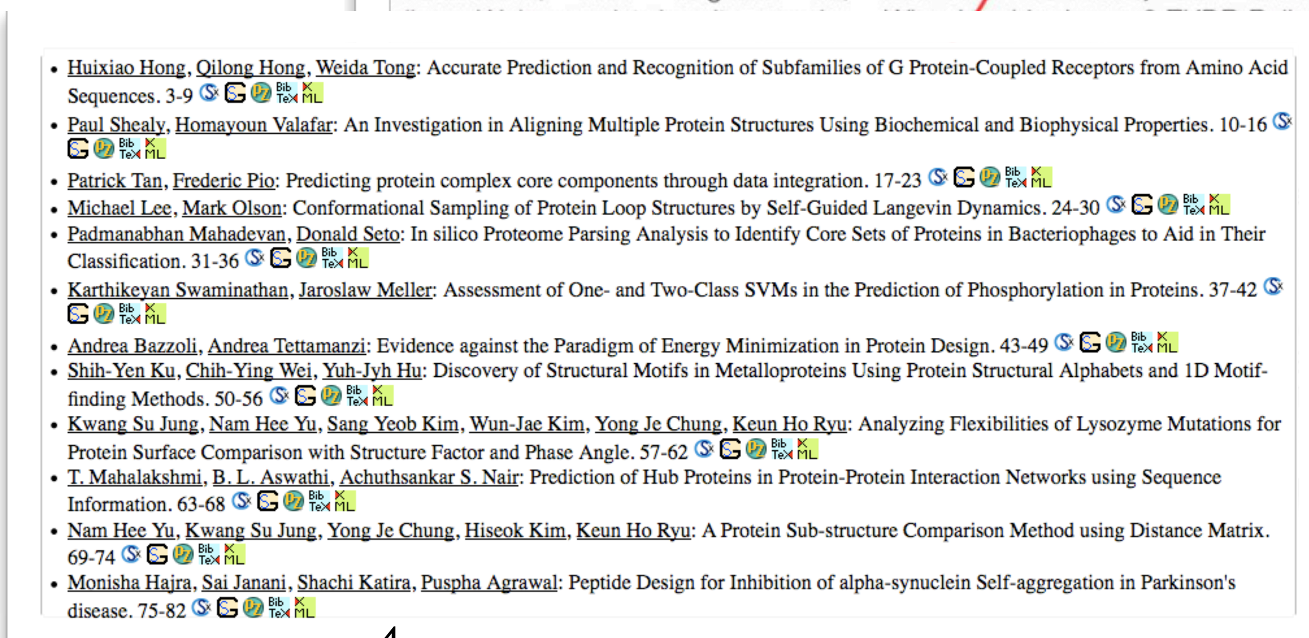
MG-Zentrum, Aachener Str. 19, 3 Etagen, 11 Zimmer, Küche, Diele, 2 Bäder, Gäste-WC, 1 Dachterrasse (192m²), 3 Balkone, 2 Enten, 3 Kaninchen, 6 Teiche, 210m² Wohnfläche, KM vergesst es,

MG Viktorieastr. 2ZKDB Blk. 55m², ren. 280 + NK + Kaut. 0176/51150216

MG-RY, Unterheydener Str. 2.Etg. 3ZKDB ca. 67m² Gas-Etg.-Hzg. an max. 3P. ab 1.10.08 KM 375,- + NK + 2MMKt. Tel.: 02166/1470598

MG, 3ZKDB, 67m², 468,50 EUR warm, ab 1.8. Tel.:02161/583234

MG-Rönneterheide, sehr schöne



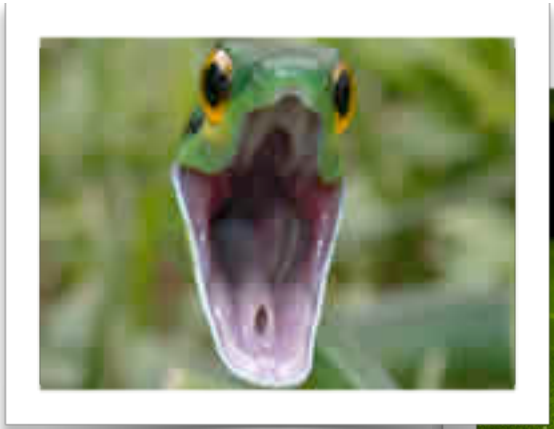
- Huixiao Hong, Qilong Hong, Weida Tong: Accurate Prediction and Recognition of Subfamilies of G Protein-Coupled Receptors from Amino Acid Sequences. 3-9
- Paul Shealy, Homayoun Valafar: An Investigation in Aligning Multiple Protein Structures Using Biochemical and Biophysical Properties. 10-16
- Patrick Tan, Frederic Pio: Predicting protein complex core components through data integration. 17-23
- Michael Lee, Mark Olson: Conformational Sampling of Protein Loop Structures by Self-Guided Langevin Dynamics. 24-30
- Padmanabhan Mahadevan, Donald Seto: In silico Proteome Parsing Analysis to Identify Core Sets of Proteins in Bacteriophages to Aid in Their Classification. 31-36
- Karthikeyan Swaminathan, Jaroslaw Meller: Assessment of One- and Two-Class SVMs in the Prediction of Phosphorylation in Proteins. 37-42
- Andrea Bazzoli, Andrea Tettamanzi: Evidence against the Paradigm of Energy Minimization in Protein Design. 43-49
- Shih-Yen Ku, Chih-Ying Wei, Yuh-Jyh Hu: Discovery of Structural Motifs in Metalloproteins Using Protein Structural Alphabets and 1D Motif-finding Methods. 50-56
- Kwang Su Jung, Nam Hee Yu, Sang Yeob Kim, Wun-Jae Kim, Yong Je Chung, Keun Ho Ryu: Analyzing Flexibilities of Lysozyme Mutations for Protein Surface Comparison with Structure Factor and Phase Angle. 57-62
- T. Mahalakshmi, B. L. Aswathi, Achuthsankar S. Nair: Prediction of Hub Proteins in Protein-Protein Interaction Networks using Sequence Information. 63-68
- Nam Hee Yu, Kwang Su Jung, Yong Je Chung, Hiseok Kim, Keun Ho Ryu: A Protein Sub-structure Comparison Method using Distance Matrix. 69-74
- Monisha Hajra, Sai Janani, Shachi Katira, Puspha Agrawal: Peptide Design for Inhibition of alpha-synuclein Self-aggregation in Parkinson's disease. 75-82

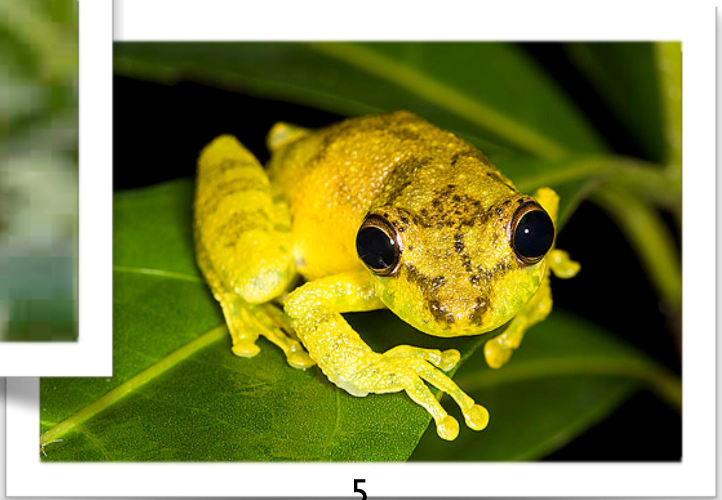
Example

```
Leptophis ahaetulla, road to Overtoom, in bush
above water in the process of eating Hyla minuta
16-V-1968. RMNH 15100

Hyla minuta 1♀ 2♂ Las Claritas, 9-VI-1978 quaking
near water 50 cm above water surface, near secondary
vegetation, 200 m, M.S. Hoogmoed, RMNH 27217 27219
```

Descriptions of two zoological specimen





Pitfalls

```
Leptophis ahaetulla, road to Overtoom, in bush
above water in the process of eating Hyla minuta
16-V-1968. RMNH 15100

Hyla minuta 1♀ 2♂ Las Claritas, 9-VI-1978 quaking
near water 50 cm above water surface, near secondary
vegetation, 200 m, M.S. Hoogmoed, RMNH 27217 27219
```

genus
species
gender
place
biotope
remark
date
collector
reg.no.

Pitfalls

```
Leptophis ahaetulla, road to Overtoom, in bush
above water in the process of eating Hyla minuta
16-V-1968. RMNH 15100

Hyla minuta 1♀ 2♂ Las Claritas, 9-VI-1978 quaking
near water 50 cm above water surface, near secondary
vegetation, 200 m, M.S. Hoogmoed, RMNH 27217 27219
```

genus
species
gender
place
biotope
remark
date
collector
reg.no.

- missing entries
- variable ordering
- mixed delimiters
- variable length
- encoding (e.g. date)

Databases

Goal: transform semi-structured text into database

Field	Entry 1	Entry 2
genus	Leptophis	Hyla
species	ahaetulla	minuta
gender	-	1 male; 2 female
place	road to Overtoom	Las Claritas
biotope	in bush above water	quaking near water 50 cm
remark	in the process of eating	-
date	16/05/1968	09/06/1978
collector	-	M.S. Hoogmoed
reg.no	15100	27217; 27219

Databases

Goal: transform semi-structured text into database

Field	Entry 1	Entry 2
genus	Leptophis	Hyla
species	ahaetulla	minuta
gender	-	1 male; 2 female
place	road to Overtoom	Las Claritas
biotope	in bush above water	quaking near water 50 cm
remark	in the process of eating	-
date	16/05/1968	09/06/1978
collector	-	M.S. Hoogmoed
reg.no	15100	27217; 27219

gain structure but implies loss of information!

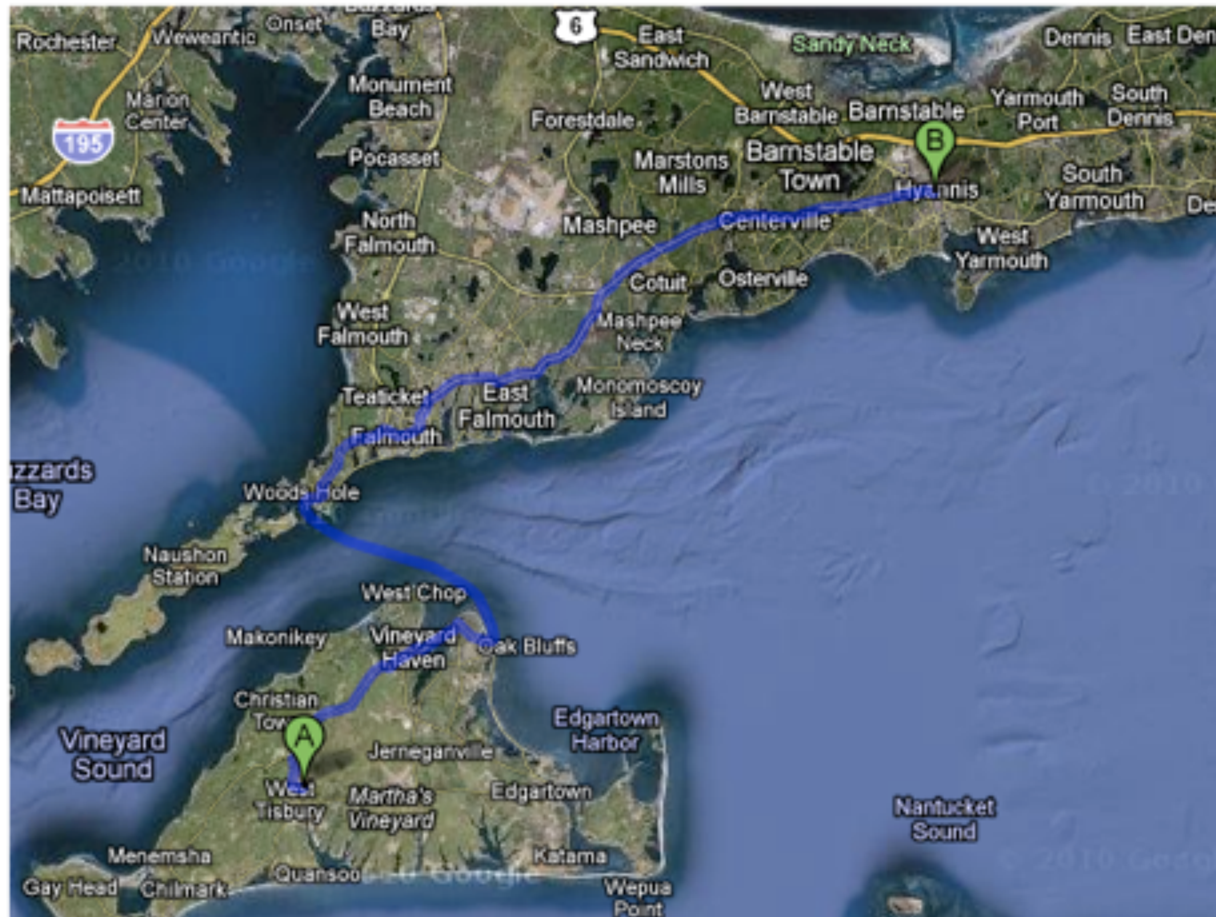
Why use Databases?

Structured text gives lots of advantages:

We can formulate complex queries over database entries

E.g. : All locations of a certain collector sorted by date => visualize by map

Citation flow graph



Why use Databases?

Structured text gives lots of advantages:

We can formulate complex queries over database entries

E.g. : All locations of a certain collector sorted by date => visualize by map

Citation flow graph

Main Question

How can we transform a semi-structured text into a database format?

Task known as: Field Segmentation

"Field segmentation refers to the automated finding and labeling in object or event descriptions"

Requirements

How can we transform a semi-structured text into a database format?

Requirements (for a good method):

- Low error rate
- Robust
- Reusable
- Unsupervised (or at least few training)

Methods

- By manual inspection: expensive, error prone, often requires domain experts
- Apply methods from CS:
  - Write a parser or rule set: not reusable, deals badly semi-structured text
  - Probabilistic methods: apply supervised or unsupervised machine learning techniques

Methods

- Almost all common machine learning methods for field segmentation are supervised
- e.g. using Hidden Markov Models or trained context free grammars.
- Drawback: Requires effort to generate training data

Methods

How to bootstrap a field segmentation algorithm from an existing database?

=> Approach by S. Canisius and C. Sporleder:

Dataset

For the evaluation of the method two datasets were used:

- RA dataset: field book about reptiles and amphibians; 16670 entries in DB; 19 fields
- Pisces dataset: field book about fish specimen; 1375 entries in DB; 4 fields

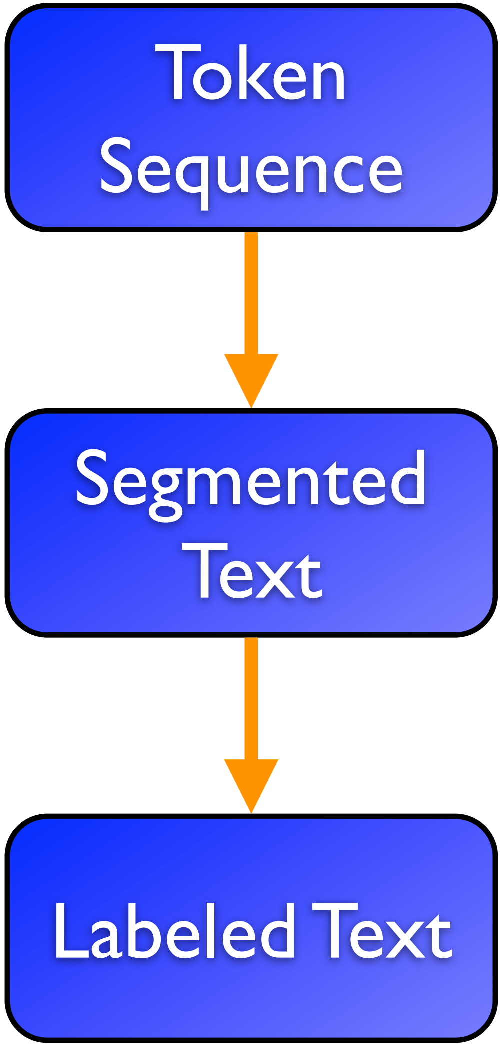
Both datasets provided by the Dutch National Museum of Natural History

Field Segmenter

Main Ideas:

- Use a trained language model to partition a semi-structured text into pre-segmentation

- A Hidden Markov Model assigns the most likely label to each segment



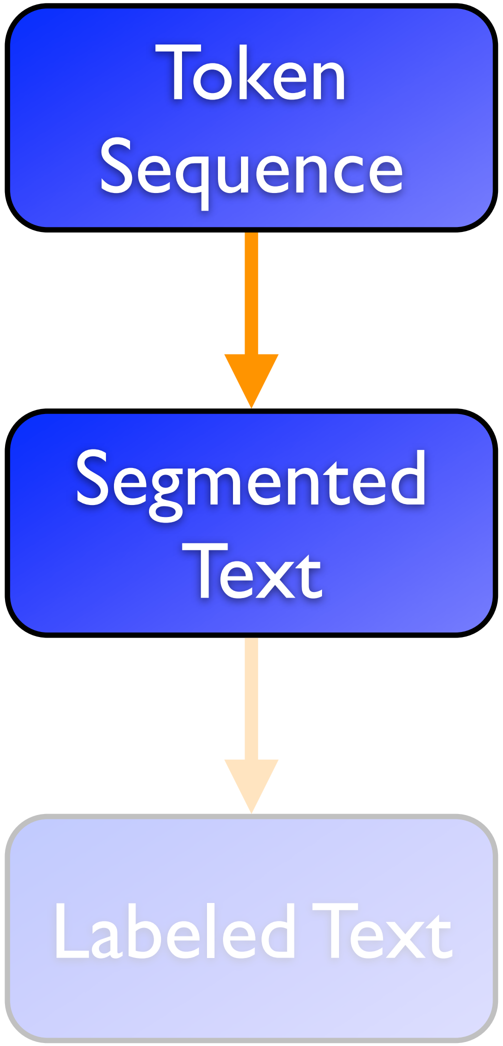
Segmentation Model

Assumption:

Segment boundaries are due to unlikely tokens

Train bigram language with entries in your database

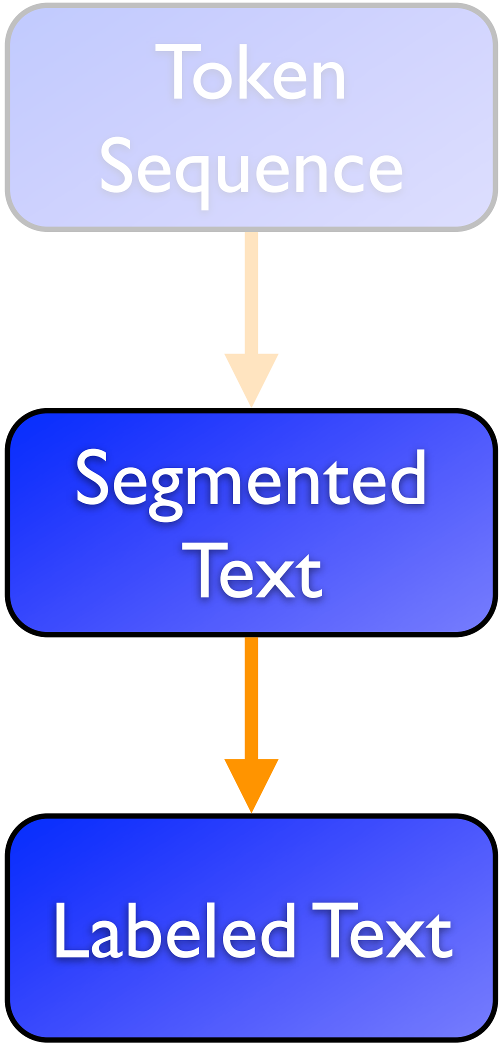
=> Use Viterbi with the language model to obtain a segmentation



HMM Parameters

For a HMM several parameters have to be derived from the data:

- Initial distribution:
  $P(X_0=s_i)$
- State-transition distribution:
  $P(X_t=s_i|X_{t-1}=s_j)$
- State-emission distribution:
  $P(O_t=o_i|X_t=s_i)$

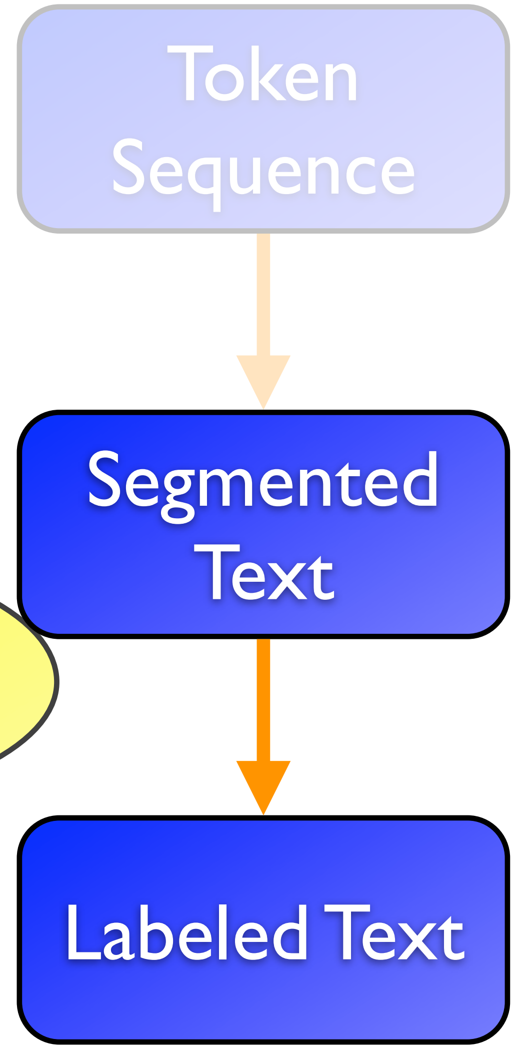


HMM Parameters

For a HMM several parameters have to be derived from the data:

- Initial distribution:
  $P(X_0=s_i)$
- State-transition distribution:
  $P(X_t=s_i|X_{t-1}=s_j)$
- State-emission distribution:
  $P(O_t=o_i|X_t=s_i)$

Use your database



HMM Parameters

For a HMM several parameters have to be derived from the data:

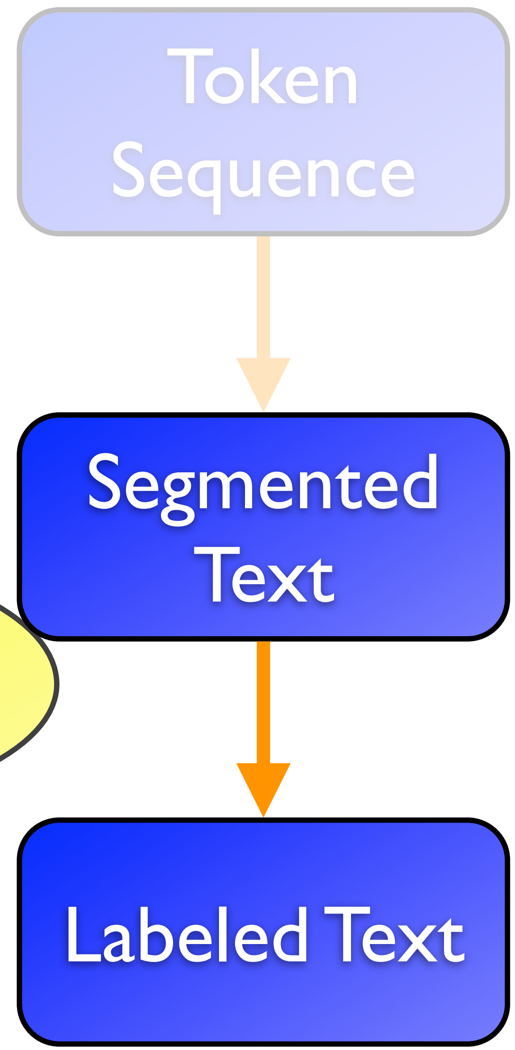
- Initial distribution:
  $P(X_0=s_i)$

- State-transition distribution:
  $P(X_t=s_i|X_{t-1}=s_j)$

Use your database

- State-emission distribution:
  $P(O_t=o_i|X_t=s_i)$

For the rest: Use Baum-Welch algorithm

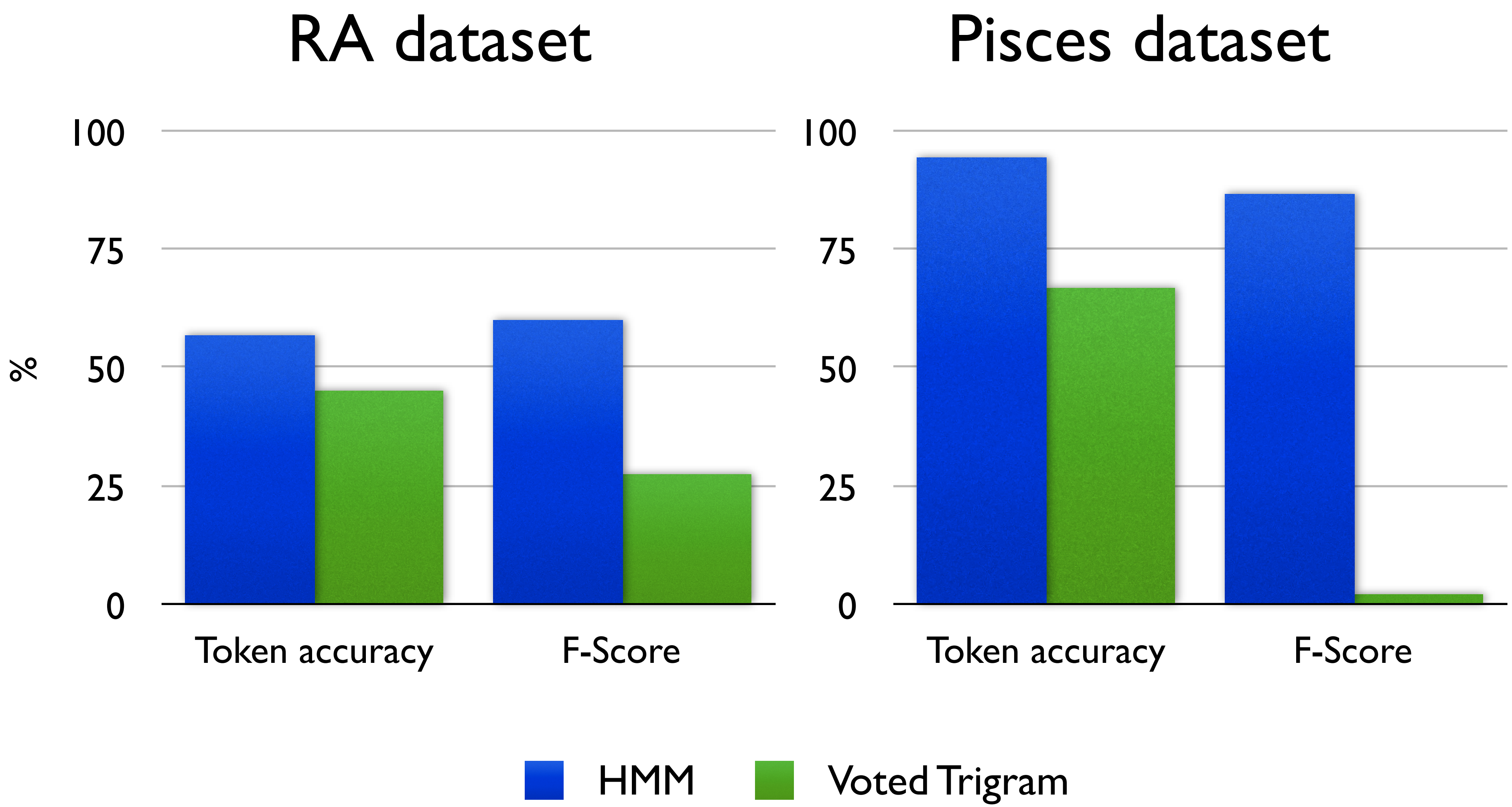


Baseline

The HMM is evaluated on RA and Pisces against several baselines:

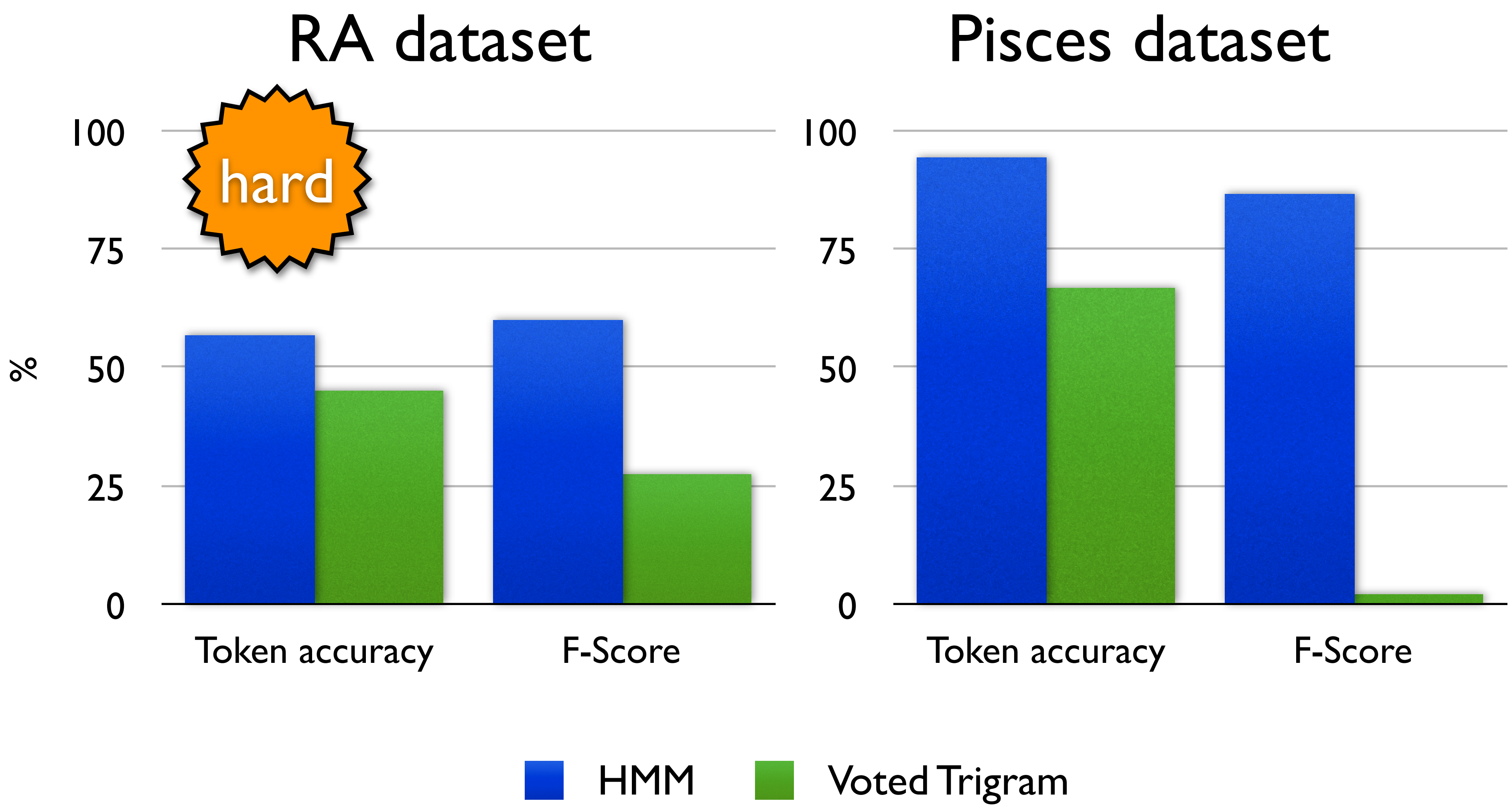
- Majority: always assign
- Exact: match longest substring with DB
- Unigram: match most likely DB entry
- Trigram: match most likely DB entry
- Voted trigram: match most likely DB entry over all trigrams

Results



RA dataset
Pisces dataset
100
75
50
25
0
%
Token accuracy
F-Score
HMM
Voted Trigram

Results



RA dataset
Pisces dataset
hard
%
100
75
50
25
0
Token accuracy
F-Score
HMM
Voted Trigram

Conclusion

- Bootstrapping a field segmenting method is possible
- You won’t get it for free, but with very few training data
- All necessary information can be derived from a preexisting database

That's it...

Thanks for your attention. Questions?