

- Gold-standard corpus
- Extract features
- Train model
- Apply to unseen data

- FLAIR web search

Crash course in Machine Learning

Teaching computers to help us teach language

Robert Reynolds

Office of Digital Humanities
Brigham Young University

CALICO
31 May 2018

Introduction

Crash course in ML

- Gold-standard corpus
- Extract features
- Train model
- Apply to unseen data

Real-world application

- FLAIR web search

- ▶ Crash course in machine learning (using example of *automatic second-language readability classification*)
 - ▶ Collect gold-standard corpus
 - ▶ Extract “features”
 - ▶ Train and evaluate a model
 - ▶ Apply model to real-world texts
- ▶ Example application of a machine-learning model
 - ▶ Web search for language learning/teaching

Collect gold-standard corpus

ML crash course

Robert Reynolds

Introduction

Crash course in
ML

- **Gold-standard
corpus**

- Extract features
- Train model
- Apply to unseen
data

Real-world
application

- FLAIR web
search

- ▶ Lots of examples of human performance of the target task
- ▶ For readability: given a text, I want to output a difficulty score
 - ▶ CEFR levels: A1, A2, B1, B2, C1, C2
 - ▶ Get examples from all levels (even distribution is best)
- ▶ “There’s no data like more data.”

- Extract features
- Train model
- Apply to unseen data

- FLAIR web search

	Total	A1	A2	B1	B2	C1	C2
CIE	145	28	57	60	—	—	—
news	50	—	—	—	—	—	50
LingQ	3481	323	653	716	832	609	348
RK	99	40	18	17	18	6	—
TORFL	168	31	36	36	26	28	11
Zlat.	746	—	66	553	127	—	—
Comb.	4689	422	830	1382	1003	643	409

Table: Distribution of documents per level for each corpus

Text complexity features

- ▶ Lexical variability
 - ▶ e.g., how often are words repeated?
- ▶ Lexical complexity
 - ▶ e.g., avg word length in letters/morphemes/syllables.
- ▶ Lexical frequency
 - ▶ based on a relevant corpus
- ▶ Morphology
 - ▶ based on automatic part-of-speech tagging
- ▶ Syntax
 - ▶ based on automatically generated sentence diagrams/trees
- ▶ etc.

Text complexity features

▶ ID	label	doc_len	avg_word_len	type-t
▶ 01	A1	106	4.7	0.5
▶ 02	A1	101	4.5	0.4
▶ 03	A2	245	5.1	0.6
▶ 04	A2	151	5.0	0.5
▶ 05	B1	230	5.3	0.7
▶ 06	B1	272	5.2	0.7
▶ 07	B2	225	5.7	0.5
▶ 08	B2	401	5.4	0.8
▶ 09	C1	643	9.4	0.6
▶ 10	C1	530	7.4	0.7
▶ 11	C2	476	8.7	0.8
▶ 12	C2	760	9.8	0.8
▶ ...				

Train and evaluate model

ML crash course

Robert Reynolds

Introduction

Crash course in ML

- Gold-standard corpus
- Extract features
- **Train model**
- Apply to unseen data

Real-world application

- FLAIR web search

- ▶ Set aside part of your data for evaluation (test/validation)
- ▶ Feed the remaining data into a machine-learning algorithm (training/development)
 - ▶ You don't have to know how these work to use them (but it helps!)
 - ▶ Your favorite programming language probably has multiple free libraries with many off-the-shelf algorithms included.
- ▶ Use the resulting model to predict labels from the test data.
- ▶ Compare the actual labels in your test data to your models predictions.

Random Forest classifier

- Gold-standard corpus
- Extract features
- **Train model**
- Apply to unseen data

- FLAIR web search

Classifier	Precision	Recall	F-score
ZeroR	0.097	0.312	0.149
OneR	0.487	0.497	0.471
RandomForest	0.690	0.677	0.671

Table: Baseline and RandomForest results with Combined corpus

Combined corpus Random Forest confusion matrix

ML crash course

Robert Reynolds

Introduction

Crash course in ML

- Gold-standard corpus
- Extract features
- **Train model**
- Apply to unseen data

Real-world application

- FLAIR web search

	A1	A2	B1	B2	C1	C2	<- classified as
A1	234	120	48	0	0	0	
A2	41	553	192	17	0	0	
B1	16	76	1130	90	5	5	
B2	1	57	311	478	83	4	
C1	1	20	66	98	394	6	
C2	0	3	40	58	9	78	

Table: Confusion matrix for RandomForest, all features, Combined corpus

Adjacent accuracy: 0.919

Use your model in real-world applications

ML crash course

Robert Reynolds

Introduction

Crash course in
ML

- Gold-standard corpus
- Extract features
- Train model
- **Apply to unseen data**

Real-world
application

- FLAIR web search

- ▶ For each new text...
 - ▶ Extract features just like you did with the gold-standard corpus
 - ▶ Your model can make a prediction based on those features

Web search for language learners and teachers

ML crash course

Robert Reynolds

Introduction

Crash course in
ML

- Gold-standard corpus
- Extract features
- Train model
- Apply to unseen data

Real-world
application

- **FLAIR web search**

- ▶ While preparing lessons and tests, how many hours have you spent looking for the perfect text?
 - ▶ right length
 - ▶ right reading level
 - ▶ target grammar topic
 - ▶ ...but NOT that other confusing grammar topic

Web search for language learners and teachers

ML crash course

Robert Reynolds

Introduction

Crash course in
ML

- Gold-standard corpus
- Extract features
- Train model
- Apply to unseen data

Real-world
application

- **FLAIR web search**

- ▶ While preparing lessons and tests, how many hours have you spent looking for the perfect text?
 - ▶ right length
 - ▶ right reading level
 - ▶ target grammar topic
 - ▶ ...but NOT that other confusing grammar topic
- ▶ <http://sifnos.sfs.uni-tuebingen.de/FLAIR/>
 - ▶ Chinkina & Meurers, 2016

The screenshot shows a web browser window with the address bar displaying "Not Secure | sifnos.sfs.uni-tuebingen.de/FLAIR/". The page has a brown header with the "FLAIR" logo and a gear icon. Below the header, there are two main sections: a "Search" section with a magnifying glass icon and a "Configure" section with a gear icon. The "Search" section contains the text: "Click on the search icon below and type in a query. FLAIR will fetch the top results from the Bing Search Engine." Below these sections, there is a white search bar with the text "Enter a query" and a magnifying glass icon. The search bar contains the text "Illini". To the right of the search bar, there is a dropdown menu showing "English" and another dropdown menu showing "10 Results". At the bottom right of the search bar, there are two buttons: "CANCEL" and "SEARCH".

FLAIR

Search

Click on the search icon below and type in a query. FLAIR will fetch the top results from the Bing Search Engine.

Configure

Enter a query

Illini

English

10 Results

CANCEL SEARCH

The screenshot shows a web browser window with the address bar displaying 'sifnos.sfs.uni-tuebingen.de/FLAIR/'. The page has an orange header with the 'FLAIR' logo and a settings gear icon. Below the header, the search term 'Illini' is centered. A red 'CANCEL ANALYSIS' button with a close icon is positioned below the search term. The main content area displays three search results, each with a green upward arrow on the left and a three-dot menu on the right. The results are as follows:

- 1** [Illinois Fighting Illini - Wikipedia](https://en.wikipedia.org/wiki/Illinois_Fighting_Illini)
The current head coach of the University of Illinois Fighting Illini Wrestling team is Jim Heffernan under his 5th season, and 22nd with the University of Illinois.
- 2** [Illini Sports | News-Gazette.com](http://www.news-gazette.com/sports/illini-sports)
Illini Sports Tipoff: The Big Ten's best 2019 prospects not yet committed. Illinois in the running for several of these blue-chippers
- 3** [Illini - Wikipedia](https://en.wikipedia.org/wiki/Illini)
Illini can refer to: Illiniwek, an alternate name for the Illinois Confederation, a group of Native American tribes in the upper Mississippi River Valley;

On the right side of the page, there are two circular buttons: an orange one with a magnifying glass icon and a blue one with an upload icon. At the bottom of the browser window, a standard navigation bar with back, forward, and search icons is visible.

The screenshot shows the FLAIR web application interface. The browser address bar displays "Not Secure | sifnos.sfs.uni-tuebingen.de/FLAIR/". The FLAIR logo is in the top left, and navigation icons are in the top right. The main content area shows search results for the query "Illini".

Search Results:

- 1 **Illini**
illiniine.com
Advanced Search : Sign Up | Forgot Password?: Home; About Us; General Info; Downloads; Quick Links; My Account
- 2 **Illinois Fighting Illini - Wikipedia**
https://en.wikipedia.org/wiki/Illinois_Fighting_Illini
The current head coach of the University of Illinois Fighting Illini Wrestling team is Jim Heffernan under his 5th season, and 22nd with the University of Illinois.
- 3 **IllinoisLoyalty.com - Illinois Fighting Illini Athlet...**
www.illinoisloyalty.com
Illinois Fighting Illini Athletics Sports Blog - IllinoisLoyalty.com. Home; Forums; Illini Basketball; Illini Football; Illini Tickets; ... 2017-18 Illini Golf
- 4 **Illini Sports | News-Gazette.com**
www.news-gazette.com/sports/illini-sports
Illini Sports Tipoff: The Big Ten's best 2019 prospects not yet

Left Sidebar:

- 10 Results (0 Filtered)
- VISUALIZE
- SHARE SEARCH SETUP
- Text Characteristics:**
 - Length:
 - ☐ Prefer shorter texts.
 - Levels:
 - ✓ A1-A2 (0 / 10)
 - ✓ B1-B2 (4 / 10)
 - ✓ C1-C2 (6 / 10)
- Constructions:**
 - ✓ Sentences
 - ✓ Parts of Speech

The screenshot displays the FLAIR web application interface. The browser address bar shows the URL `sifnos.sfs.uni-tuebingen.de/FLAIR/`. The application header is orange with the FLAIR logo and navigation icons. The main content area shows 10 search results for the query 'Illini'. On the left, there are two sidebars: 'Text Characteristics' and 'Constructions'. The 'Text Characteristics' sidebar includes a 'Length' section with a checked option 'Prefer shorter texts.' and a 'Levels' section with checked options 'A1-A2', 'B1-B2', and 'C1-C2'. The 'Constructions' sidebar has a checked option 'Sentences' and an unchecked option 'Parts of Speech'. The search results list includes:

- 1. **Illini** (illini.com) - Advanced Search : Sign Up | Forgot Password?: Home; About Us; General Info; Downloads; Quick Links; My Account
- 2. **Illinois Fighting Illini College Basketball - ESPN...** (www.espn.com/mens-college-basketball/team/_id/356) - Get the latest Illinois Fighting Illini news, scores, stats, standings, rumors, and more from ESPN.
- 3. **Illini - Wikipedia** (https://en.wikipedia.org/wiki/Illini) - Illini can refer to: Illiniwek, an alternate name for the Illinois Confederation, a group of Native American tribes in the upper Mississippi River Valley;
- 4. **Writing Illini - An Illinois Fighting Illini Fan Site ...** (https://writingillini.com) - The ultimate home for Illinois Fighting Illini news, rumors, player and team updates, commentary, recruiting, analysis, and

At the bottom of the interface, there is a navigation bar with various icons for document manipulation and search.

FLAIR

Not Secure | sifnos.sfs.uni-tuebingen.de/FLAIR/

← FLAIR

4 Results (6 Filtered)

VISUALIZE

SHARE SEARCH SETUP

Text Characteristics:

Length:

- ✓ Prefer shorter texts.

Levels:

- ✓ A1-A2 0 / 4
- ✓ B1-B2 4 / 4
- ☐ C1-C2 0 / 4

Constructions:

- ✓ Sentences
- ✓ Parts of Speech

'Illini'

- Illini - Wikipedia**
<https://en.wikipedia.org/wiki/Illini>
Illini can refer to: Illiniwek, an alternate name for the Illinois Confederation, a group of Native American tribes in the upper Mississippi River Valley;
- Writing Illini - An Illinois Fighting Illini Fan Site ...**
<https://writingillini.com>
The ultimate home for Illinois Fighting Illini news, rumors, player and team updates, commentary, recruiting, analysis, and opinion. Covering Illinois football, Illinois basketball, Illinois baseball, and more!
- Illini Sports | News-Gazette.com**
www.news-gazette.com/sports/illini-sports
Illini Sports Tipoff: The Big Ten's best 2019 prospects not yet committed. Illinois in the running for several of these blue-chippers
- Illinois Fighting Illini - Wikipedia**
https://en.wikipedia.org/wiki/Illinois_Fighting_Illini

The screenshot shows the FLAIR web application interface. The browser address bar displays the URL `sifnos.sfs.uni-tuebingen.de/FLAIR/`. The application has an orange header bar with the FLAIR logo and navigation icons. On the left, a sidebar contains a 'Sentence Types' section with four options: Simple, Coordinate, Subordinate, and Incomplete, each with a slider and a '4/4' indicator. Below this is a 'Clause Types' section. The main content area displays search results for the query 'Illini'. The results are numbered 1 through 4, each with a green upward arrow icon, a title, a URL, and a brief description. The results are: 1. Illini - Wikipedia (https://en.wikipedia.org/wiki/Illini), 2. Writing Illini - An Illinois Fighting Illini Fan Site ... (https://writingillini.com), 3. Illini Sports | News-Gazette.com (www.news-gazette.com/sports/illini-sports), and 4. Illinois Fighting Illini - Wikipedia (https://en.wikipedia.org/wiki/Illinois_Fighting_Illini). On the right side of the main content area, there are two circular buttons: a magnifying glass icon and an upload icon.

FLAIR

Sentence Types

- Simple 4/4
- Coordinate 4/4
- Subordinate 4/4
- Incomplete 4/4

Clause Types

'Illini'

- Illini - Wikipedia**
<https://en.wikipedia.org/wiki/Illini>
Illini can refer to: Illiniwek, an alternate name for the Illinois Confederation, a group of Native American tribes in the upper Mississippi River Valley;
- Writing Illini - An Illinois Fighting Illini Fan Site ...**
<https://writingillini.com>
The ultimate home for Illinois Fighting Illini news, rumors, player and team updates, commentary, recruiting, analysis, and opinion. Covering Illinois football, Illinois basketball, Illinois baseball, and more!
- Illini Sports | News-Gazette.com**
www.news-gazette.com/sports/illini-sports
Illini Sports Tipoff: The Big Ten's best 2019 prospects not yet committed. Illinois in the running for several of these blue-chippers
- Illinois Fighting Illini - Wikipedia**
https://en.wikipedia.org/wiki/Illinois_Fighting_Illini

The screenshot shows the FLAIR web interface. The browser address bar displays "Not Secure | sifnos.sfs.uni-tuebingen.de/FLAIR/". The FLAIR logo is in the top left, and navigation icons are in the top right. On the left, a sidebar contains a "Sentence Types" section with four filters: "Simple" (checked, 1/4), "Coordinate" (checked, 4/4), "Subordinate" (checked, 4/4), and "Incomplete" (checked, 4/4). Each filter has a blue slider. Below this is a "Clause Types" section. The main content area displays search results for the query "'Illini'". The results are numbered 1 through 4. Result 1 is "Illinois Fighting Illini - Wikipedia" with a URL and a snippet about the head coach. Result 2 is "Illini Sports | News-Gazette.com" with a URL and a snippet about the Big Ten's best 2019 prospects. Result 3 is "Writing Illini - An Illinois Fighting Illini Fan Site ..." with a URL and a snippet about the fan site. Result 4 is "Illini - Wikipedia" with a URL. On the right side of the interface, there are two circular buttons: a magnifying glass for search and an upload icon.

FLAIR

Sentence Types

- ✓ Simple 1/4
- ✓ Coordinate 4/4
- ✓ Subordinate 4/4
- ✓ Incomplete 4/4

Clause Types

'Illini'

- Illinois Fighting Illini - Wikipedia**
https://en.wikipedia.org/wiki/Illinois_Fighting_Illini
The current head coach of the University of Illinois Fighting Illini Wrestling team is Jim Heffernan under his 5th season, and 22nd with the University of Illinois.
- Illini Sports | News-Gazette.com**
www.news-gazette.com/sports/illini-sports
Illini Sports Tipoff: The Big Ten's best 2019 prospects not yet committed. Illinois in the running for several of these blue-chippers
- Writing Illini - An Illinois Fighting Illini Fan Site ...**
<https://writingillini.com>
The ultimate home for Illinois Fighting Illini news, rumors, player and team updates, commentary, recruiting, analysis, and opinion. Covering Illinois football, Illinois basketball, Illinois baseball, and more!
- Illini - Wikipedia**
<https://en.wikipedia.org/wiki/Illini>

Today's panel

ML crash course

Robert Reynolds

Introduction

Crash course in
ML

- Gold-standard corpus
- Extract features
- Train model
- Apply to unseen data

Real-world
application

- FLAIR web search

- ▶ **Rob Reynolds:** readability of Russian authentic texts
- ▶ **Sowmya Vajjala:** readability of English authentic texts
- ▶ **Elena Cotos:** automated writing evaluation
- ▶ **Haiyang Ai:** correcting verb-noun miscollocations