AIDA-light: High-Throughput Named-Entity Disambiguation

Fourat Belhaj Rhouma
fouratbhr@gmail,com

Outline

Named-Entity Disambiguation (NED)

- State of the Art NED Systems
- AIDA-Light

Overview of AIDA-Light

- System Architecture

Experiment

- Discussion

Named-Entity Disambiguation

NED aims to map mentions of ambiguous names in natural language onto a set of known entities

Text & Mentions

Under Fergie, United won the Premier League title 13 times.

correct entities

Fergie_(singer), an American singer, songwriter, fashion designer, television host and actress.
Alex_Ferguson, a former Scottish football manager of Manchester United F.C.
Sarah,_Duchess_of_York, the former wife of Prince Andrew, Duke of York.
. . .

United_Airlines, an American major airline.
United_Airways, a Bangladeshi airline.
Manchester_United_F.C., an English professional football club.
. . .

Premier_League, the English professional football league.
. . .

State of the Art NED Systems

- **High-Accuracy Systems :**
  - **AIDA : use rich contextual features (and joint inference) -> emphasis on quality**

- **Fast (High-performance) Systems :**
  - **TagMe2, DBPedia Spotlight : mention-by-mention inference with more lightweight features -> emphasis on speed**

AIDA-LIGHT

- Reduce the memory used while running which contributes to faster processing
- Use a thematic domain hierarchy to capture measures for domain-entity and entity-entity coherence
- Employs a novel two-stage algorithm in which we determine the "easy and low-cost" mappings first. By doing so, it reduce both the complexity and difficulty of the second-stage disambiguations
- is a complete system for NED, which is orders of magnitude faster than AIDA while achieving comparable output quality

Outline

Named-Entity Disambiguation (NED)

- State of the Art NED Systems
- AIDA-Light

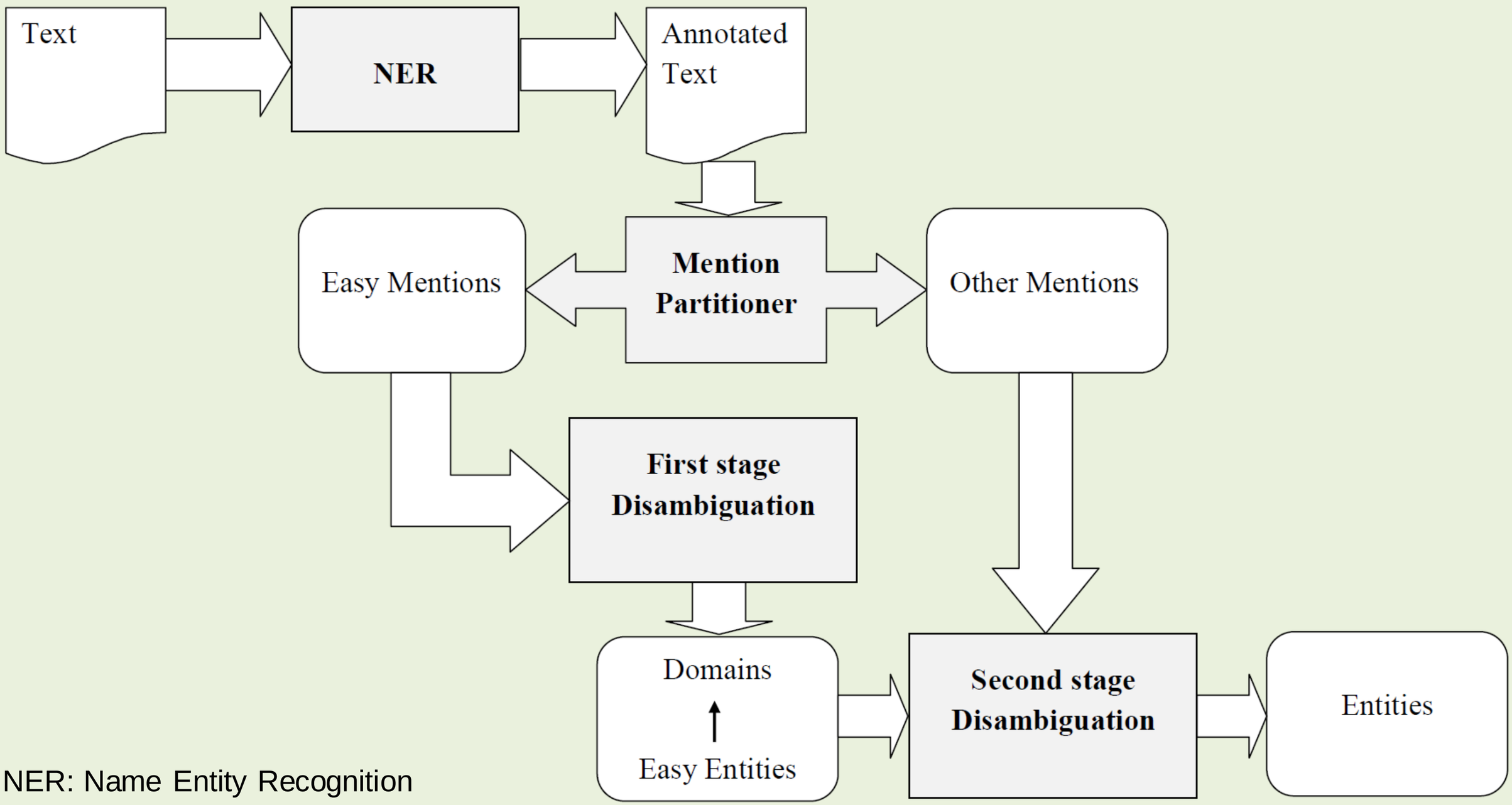
Overview of AIDA-Light

- System Architecture

Experiment

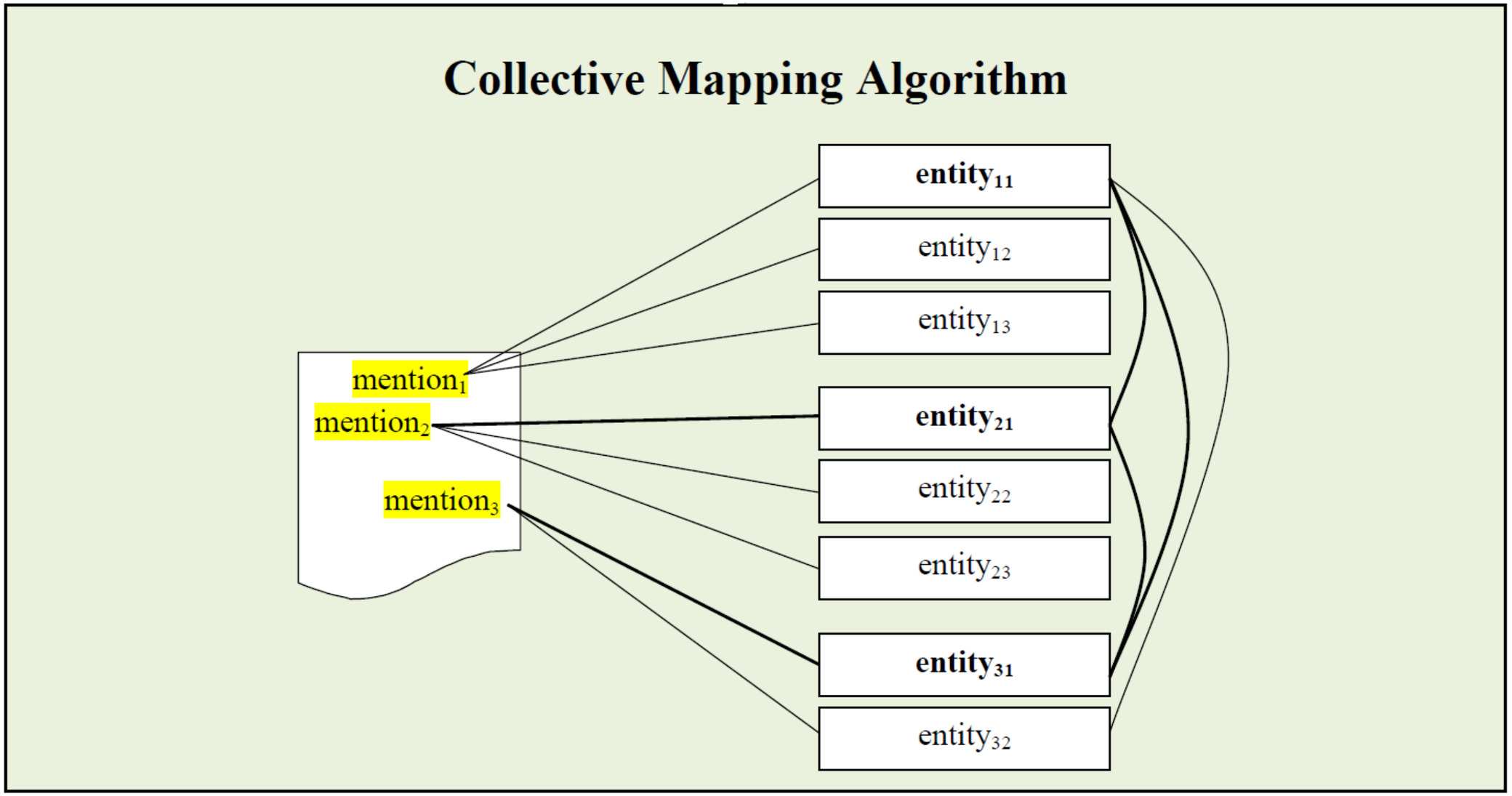
- Discussion

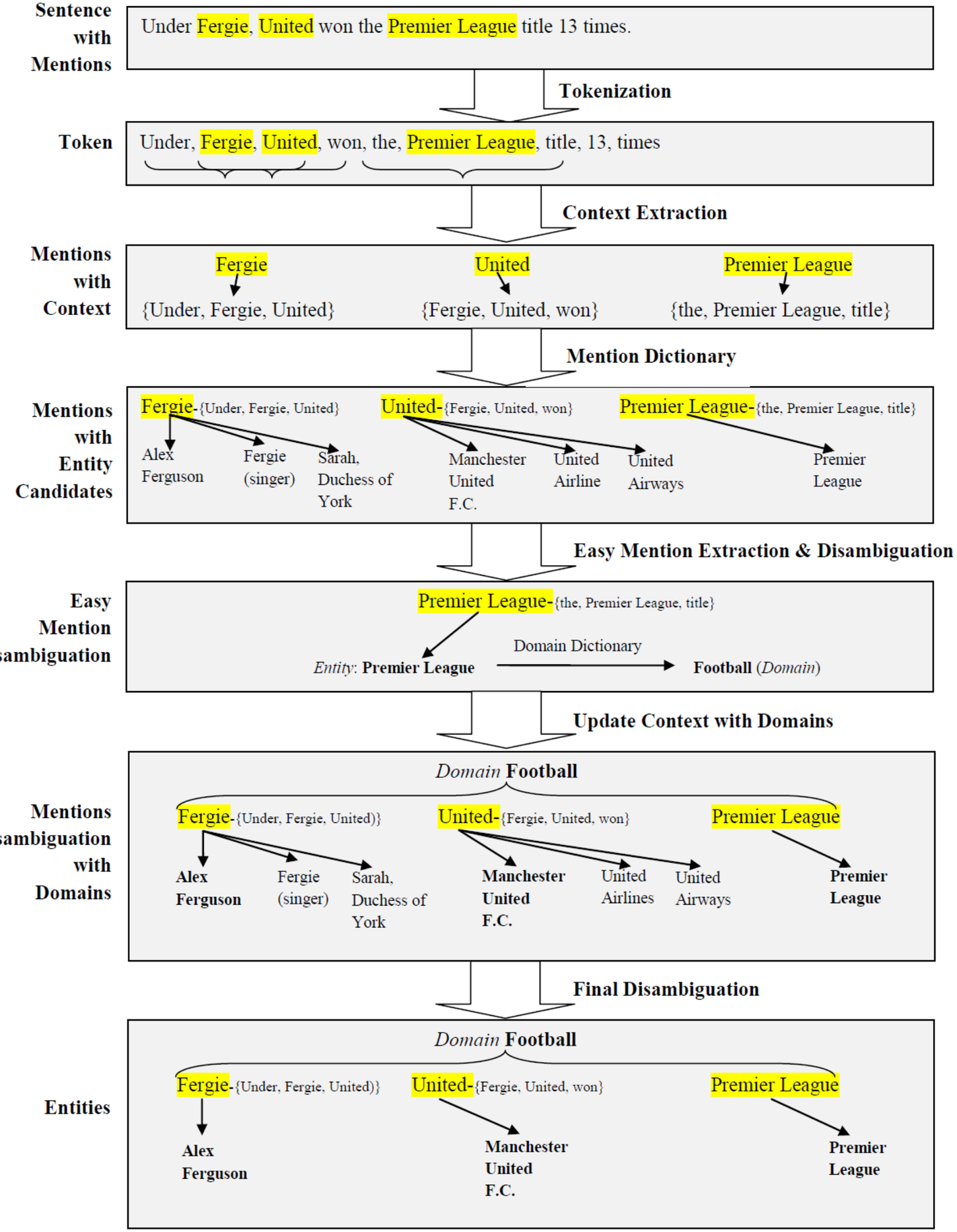
System Architecture



NER: Name Entity Recognition

2nd/Final stage Disambiguation




Sentence with Mentions
Under Fergie, United won the Premier League title 13 times.
Tokenization
Token
Under, Fergie, United, won, the, Premier League, title, 13, times
Context Extraction
Mentions with Context
Fergie
United
Premier League
{Under, Fergie, United}
{Fergie, United, won}
{the, Premier League, title}
Mention Dictionary
Mentions with Entity Candidates
Fergie-{Under, Fergie, United}
United-{Fergie, United, won}
Premier League-{the, Premier League, title}
Alex Ferguson
Fergie (singer)
Sarah, Duchess of York
Manchester United F.C.
United Airline
United Airways
Premier League
Easy Mention Extraction & Disambiguation
Easy Mention Disambiguation
Premier League-{the, Premier League, title}
Entity: Premier League
Domain Dictionary
Football (Domain)
Update Context with Domains
Mentions Disambiguation with Domains
Domain Football
Fergie-{Under, Fergie, United)}
United-{Fergie, United, won}
Premier League
Alex Ferguson
Fergie (singer)
Sarah, Duchess of York
Manchester United F.C.
United Airlines
United Airways
Premier League
Final Disambiguation
Entities
Domain Football
Fergie-{Under, Fergie, United)}
United-{Fergie, United, won}
Premier League
Alex Ferguson
Manchester United F.C.
Premier League

Outline

Named-Entity Disambiguation (NED)

- State of the Art NED Systems
- AIDA-Light

Overview of AIDA-Light

- System Architecture

Experiment

- Discussion

Experimental Corpora

- **CoNLL-YAGO testb:** news articles with long-tail entities
- **WP:** short contexts with highly ambiguous mentions and long-tail entities
- **Wikipedia articles:** Wikipedia articles with internal links as mentions
- **Wiki-links:** long documents with a few mentions

Results on NED Quality

- **Precision on different corpora, statistically significant improvements over Spotlight are marked with an asterisk**

Dataset	AIDA	AIDA-*light*	Spotlight
CoNLL-YAGO	82.5%*	**84.8%***	75.0%
WP	**84.7%***	84.4%*	63.8%
Wikipedia articles	**90.0%**	88.3%	89.6%
Wiki-links	80.3%	**85.1%***	80.7%

Results on Run-time

- **Average per-document run-time results**

Dataset	AIDA-*light*	Spotlight
CoNLL-YAGO	0.47s	0.51s
WP	0.05s	0.14s
Wikipedia articles	5.47s	4.22s
Wiki-links	0.18s	0.32s

- **AIDA uses a SQL database, not considered here**

Discussion

Cross-language search : The case of Google Language Tools

Fourat Belhaj Rhouma
fouratbhr@gmail,com

Outline

Motivation

Google Language Tools (GLT)

Google's translation mechanism

Implication of GLT

- Discussion

Motivation

- 1. To help Web users to appropriately evaluate and use GLT and other language support services on the Web.
  - By aiming to provide advice to Web users on how to appropriately use language support services through the analysis of Google Language Tools.
- 2. To identify important issues as related to technological transfer in MT and CLIR.
  - In order to transfer research achievements in CLIR and MT to practical applications.

- **GLT:** Google Language Tools
- **CLIR:** Cross-Language Information Retrieval
- **MT:** Machine Translation

Outline

Motivation

Google Language Tools (GLT)

- Cross-language search
- Other language support services

Google's translation mechanism

Implication of GLT

- Discussion

Cross-language search

A cross-language search system consists of following components:

1. **Search interface: type in search terms and specify language for these terms and for the retrieved Web pages.**
2. **Query translation: translate the user's query into the Web pages language, so the matching can be conducted**
3. **Web search or information retrieval: the actual search for relevent pages**
4. **Machine translation of results: translate the Web pages into the language of the query**
5. **Result interface: present the result to the user, result may be in their original form simultaneously**

Other language support services

- Monolingual search in your preferred language or country
- Machine translation of Web pages: some text translation, or Web page translation by typing in the Web page's URL
- User's feedback on translation: allow users to provide their own translation
- Online Dictionary lookup
- Other services:
  - User's homepage convertion
  - the option of adding a list of buttons of languages that can be added to the browser’s toolbar
  - free, downloadable toolbar that can translate any words from English to other languages

Outline

Motivation

Google Language Tools (GLT)

- Cross-language search
- Other language support services

Google's translation mechanism

Implication of GLT

- Discussion

Translation performance

- **Query in TITLE group are similar to queries a Web user submits to a search engine**
- **Queries in the DESC group are basically sentences that are similar to those constituting Web pages**

Table 2: Results of GLT's translation evaluation.

	Queries in TITLE group		Queries in DESC group	
	Google	SYSTRAN	Google	SYSTRAN
Correct	38 (76%)	26 (52%)	9 (18%)	13 (26%)
Other	12 (24%)	24 (48%)	41 (82%)	37 (74%)
Total queries	50	50	50	50

Outline

Motivation

Google Language Tools (GLT)

Google's translation mechanism

Implication of GLT

- Discussion

User's opinion of Google's Cross-language search

- **users thought that Google was doing the right thing**
- **The cross–language search is interesting and especially useful for non–English speakers**
- **Even though a given translation is not perfect, this tool would certainly help many access information in various languages**

- **machine translation (MT) was not mature enough for practical use**
- **Google poorly translated Arabic**

- **the majority of postings welcomed Google's launch of the cross–language search service with a concern over the quality of machine translation.**

Implications of GLT for Web users

- **bilingual users who formulate their queries in their native languages to retrieve documents in their second languages**
- **monolingual users who are interested in finding information in other languages**
  - **Immigrants**
  - **Investors interested in examining new markets**
  - **College students learning a foreign language**
  - **Patients or caregivers can search and find medical or treatment information from other countries or in other languages**
  - **travelers can search for local information and events using GLT en route**

Implications of GLT's cross-language search for research and developement in MT an CLIR

- **The launch of Google's cross–language search marked the first step towards practical use of CLIR in search engines**
- **It is promising that other kinds of information systems could provide cross–language search functions for their users**
- **Solutions to the problems of assisting users to understand search results should be explored in order to apply CLIR to empirical systems**
- **Better understanding of the needs of users could help in this investigation**

Implications of GLT to other information systems such as digital libraries

- **large amount of information stored in digital libraries is not accessible in a variety of search engines**
- **It would benefit global users if digital libraries were to offer multilingual information access services**

Discussion