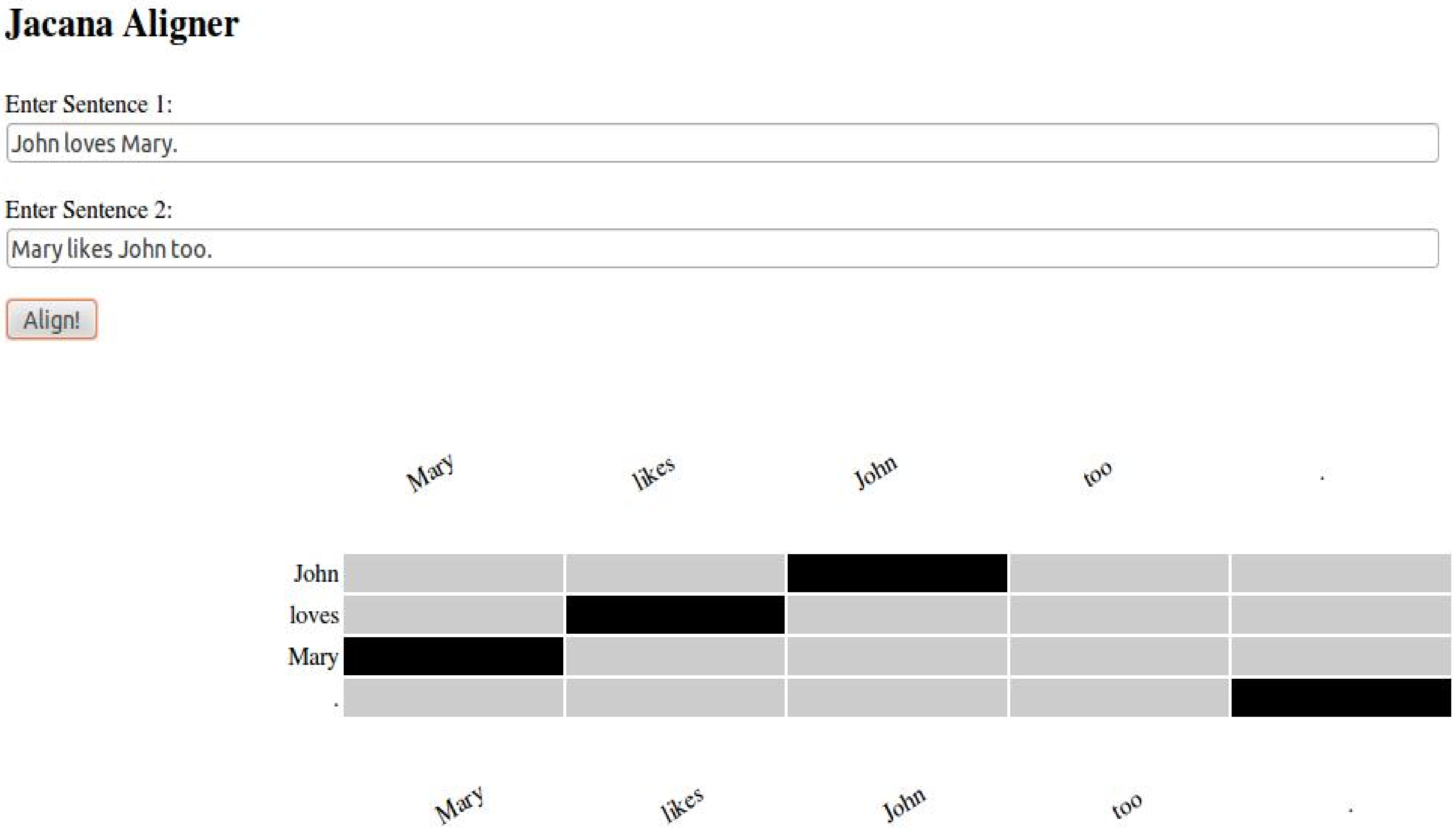
Jacana Word Aligner

Xuchen Yao

Johns Hopkins University

http://code.google.com/p/jacana

original monolingual aligner



Jacana Aligner

Enter Sentence 1:

John loves Mary.

Enter Sentence 2:

Mary likes John too.

	Mary	likes	John	too	.
John			■		
loves		■			
Mary	■				
.					■
	Mary	likes	John	too	.

http://code.google.com/p/jacana

Objective

- Extend it to MT word alignment
  - – focus on language:
    - need labeled word alignment data (German<->English, Chinese<->English, etc)
    - you know the source and target languages
    - write some simple feature functions
  - – focus on programming:
    - incorporate jwktl, the java interface to wiktionary
    - incorporate some version of “EuroWordNet”

http://code.google.com/p/jacana

Objective

- Extend it to MT word alignment
  - focus on language:
    - need labeled word alignment data (German<->English, Chinese<->English, etc)
    - you know the source and target languages
    - write some simple feature functions
  - focus on programming:
    - incorporate jwktl, the java interface to wiktionary
    - incorporate some version of “EuroWordNet”

Preliminary Results

- Used a standard French-English alignment dataset
  - state-of-the-art AER (alignment error rate) is 7%
  - I'm getting 18%, with only features based on dictionary and string similarities
  - AER should get below 10% when:
    - using more language-dependent features
    - draw stats from parallel French-English corpus

http://code.google.com/p/jacana

Preliminary Results

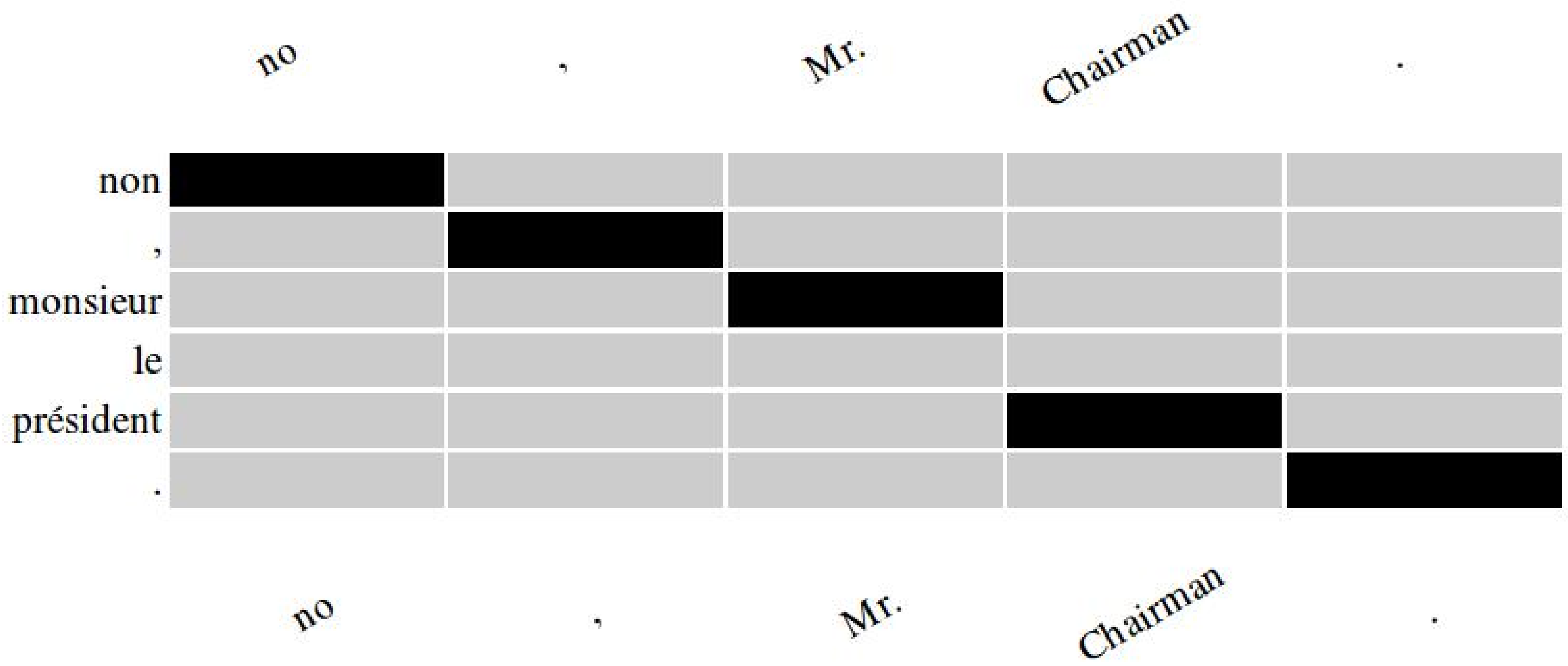
- Used a standard French-English alignment dataset
  - state-of-the-art AER (alignment error rate) is 7%
  - I'm getting 18%, with only features based on dictionary and string similarities
  - AER should get below 10% when:
    - using more language-dependent features
    - draw stats from parallel French-English corpus

http://code.google.com/p/jacana

Preliminary Results

- Used a standard French-English alignment dataset
  - state-of-the-art AER (alignment error rate) is 7%
  - I'm getting 18%, with only features based on dictionary and string similarities
  - AER should get below 10% when:
    - using more language-dependent features
    - draw stats from parallel French-English corpus

Example Output



	no	,	Mr.	Chairman	.
non	■				
,		■			
monsieur			■		
le					
président				■	
.					■
	no	,	Mr.	Chairman	.

http://code.google.com/p/jacana

Work in progress

- source code online:
  - http://code.google.com/p/jacana-xy/

- working on integration with the Joshua decoder

Give it a try!

google "jacana align"

http://code.google.com/p/jacana