CSC 439/539
Statistical Natural Language Processing
Lecture 1: Introduction

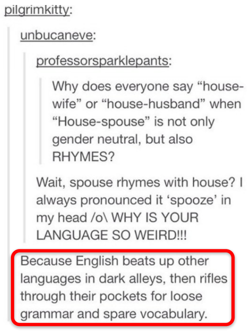
Mihai Surdeanu

Fall 2017

Take-away

- Why you should take this course
- Admin issues
- First homework due in 1 week!
- What topics will be covered in this class?

Language is hard…



pilgrimkitty:

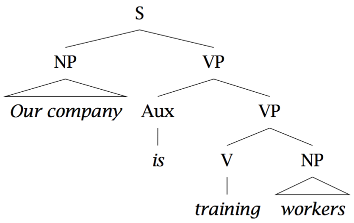
> unbucaneve:
>
> > professorsparklepants:
> >
> > > Why does everyone say "house-wife" or "house-husband" when "House-spouse" is not only gender neutral, but also RHYMES?
> >
> > Wait, spouse rhymes with house? I always pronounced it 'spooze' in my head /o\ WHY IS YOUR LANGUAGE SO WEIRD!!!
>
> Because English beats up other languages in dark alleys, then rifles through their pockets for loose grammar and spare vocabulary.

“Beating up” other languages

- Why do we eat “pork” and “beef” but we raise “pigs” and “cows”?
- What is the percentage of cognates with French in English?

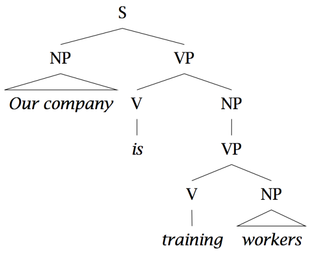
Who did what to whom?

“Our company is training workers.”



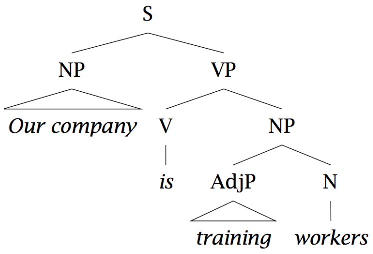
Correct: “is training” as a verb group

Who did what to whom?



Incorrect: “training” as gerund, as in:
“Our problem is training workers.”

Who did what to whom?



Incorrect: “training” modifies “workers,
as in: “Those are training wheels.”

Ambiguity and selectional preferences

I **swallowed** a bug while running.

What selectional preferences would you add for the verb “swallow”?

I **swallowed** his story, hook, line, and sinker.

The supernova **swallowed** the planet.

Variability

he acquired it

he purchased it

he bought it

it was bought by him

it was sold to him

she sold it to him

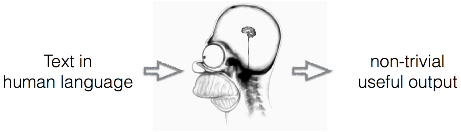
she sold him that

Slide by Yoav Goldberg

Discourse/Ellipsis/Multi-modality

BUT LANGUAGE UNDERSTANDING ENABLES IMPORTANT APPLICATIONS

NLP in a nutshell



Text in human language → non-trivial useful output

takes as input text in human language
and process it in a way that suggests
an intelligent process was involved

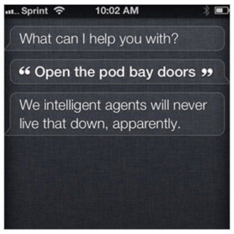
Slide by Yoav Goldberg

NLP Applications

Question Answering

NLP Applications

Question Answering



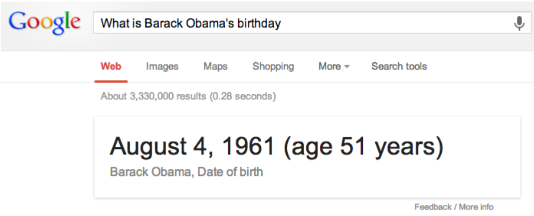
What can I help you with?

“Open the pod bay doors”

We intelligent agents will never live that down, apparently.

NLP Applications

Question Answering



What is Barack Obama's birthday

About 3,330,000 results (0.28 seconds)

August 4, 1961 (age 51 years)

Barack Obama, Date of birth

NLP Applications

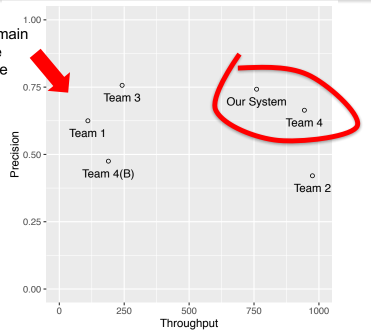
Question Answering

- When athletes begin to exercise, their heart rates and respiration rates increase. At what level of organization does the human body coordinate these functions?
  - A: at the tissue level
  - B: at the organ level
  - C: at the system level
  - D: at the cellular level

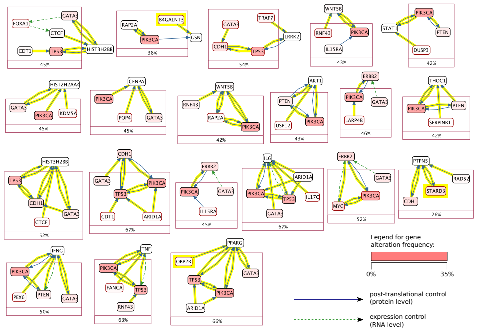
Unsolved problem!
- Needs inference
- Very little training data

Machine reading/Information extraction

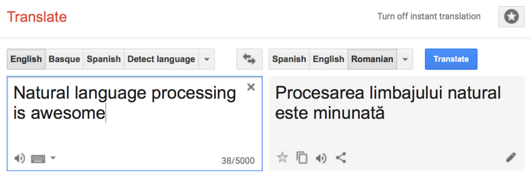
Human domain experts are around here



Machine reading/Information extraction



Machine translation



Natural language processing is awesome

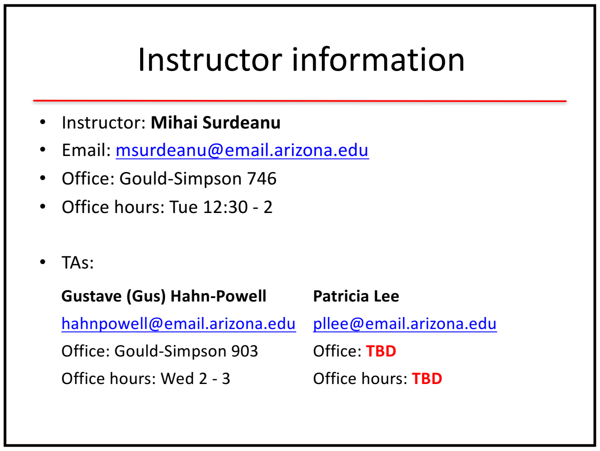
Procesarea limbajului natural este minunată

And many others…

- Can you suggest a few other NLP applications?

Overview

- Administration
- First homework
- Course overview

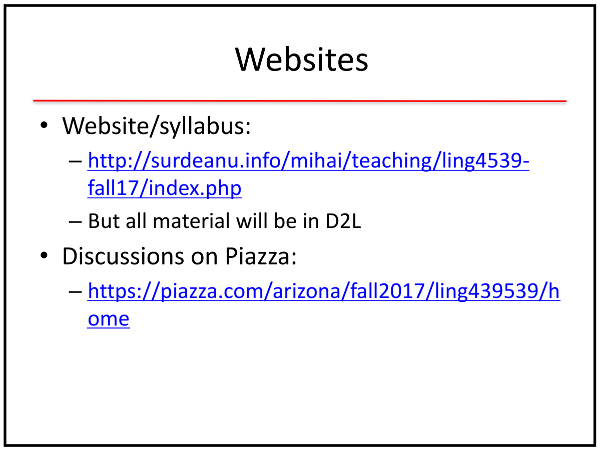


Instructor information

- Instructor: **Mihai Surdeanu**
- Email: msurdeanu@email.arizona.edu
- Office: Gould-Simpson 746
- Office hours: Tue 12:30 - 2

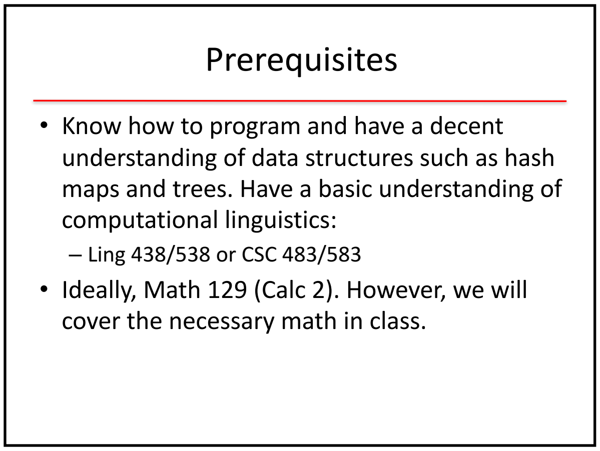
- TAs:

Gustave (Gus) Hahn-Powell	**Patricia Lee**
hahnpowell@email.arizona.edu	pllee@email.arizona.edu
Office: Gould-Simpson 903	Office: **TBD**
Office hours: Wed 2 - 3	Office hours: **TBD**



Websites

- Website/syllabus:
  - http://surdeanu.info/mihai/teaching/ling4539-fall17/index.php
  - But all material will be in D2L
- Discussions on Piazza:
  - https://piazza.com/arizona/fall2017/ling439539/home



Prerequisites

- Know how to program and have a decent understanding of data structures such as hash maps and trees. Have a basic understanding of computational linguistics:
  - Ling 438/538 or CSC 483/583
- Ideally, Math 129 (Calc 2). However, we will cover the necessary math in class.

Prerequisites: does this look scary?

```
comment: Categorization Decision
funct decision($\vec{x}, \vec{w}, \theta$) ≡
  if $\vec{w} \cdot \vec{x} > \theta$ then
            return yes
        else
            return no
  fi.
comment: Initialization
$\vec{w} = 0$
$\theta = 0$
comment: Perceptron Learning Algorithm
while not converged yet do
    for all elements $\vec{x}_j$ in the training set do
        $d$ = decision($\vec{x}_j, \vec{w}, \theta$)
        if class($\vec{x}_j$) = $d$ then
                    continue
        elsif class($\vec{x}_j$) = yes and $d$ = no then
                              $\theta = \theta - 1$
                              $\vec{w} = \vec{w} + \vec{x}_j$
        elsif class($\vec{x}_j$) = no and $d$ = yes then
                              $\theta = \theta + 1$
                              $\vec{w} = \vec{w} - \vec{x}_j$
        fi
    end
end
```

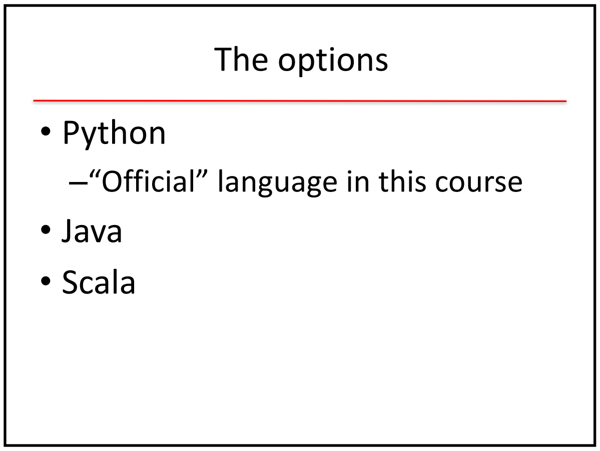
Prerequisites: does this look scary?

$$||x||_2 = \sqrt{\sum_i x_i^2}$$

$$\cos(\vec{q}, \vec{d}) = \textsc{sim}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

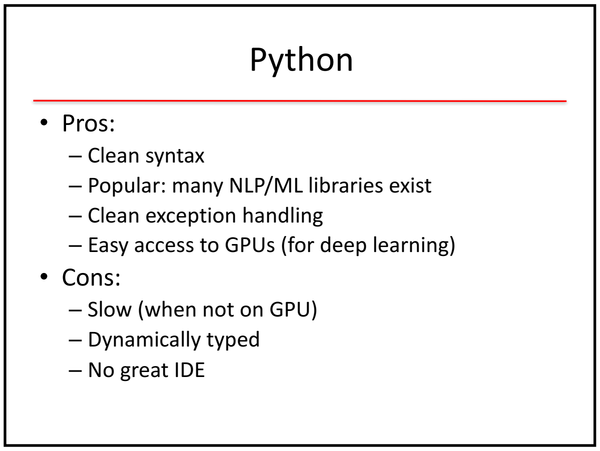
Dot product, matrix multiplication, Bayes rule

Choosing a programming language



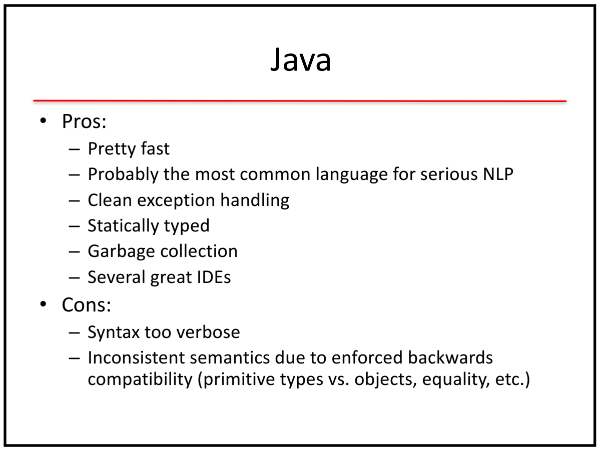
The options

- Python
  - “Official” language in this course
- Java
- Scala



Python

- Pros:
  - Clean syntax
  - Popular: many NLP/ML libraries exist
  - Clean exception handling
  - Easy access to GPUs (for deep learning)
- Cons:
  - Slow (when not on GPU)
  - Dynamically typed
  - No great IDE



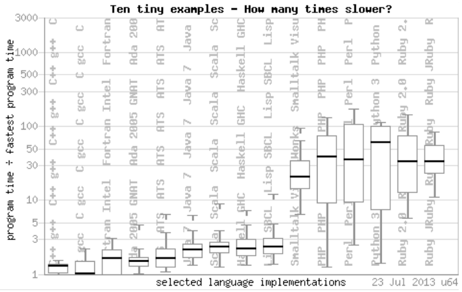
Java

- Pros:
  - Pretty fast
  - Probably the most common language for serious NLP
  - Clean exception handling
  - Statically typed
  - Garbage collection
  - Several great IDEs
- Cons:
  - Syntax too verbose
  - Inconsistent semantics due to enforced backwards compatibility (primitive types vs. objects, equality, etc.)

Scala

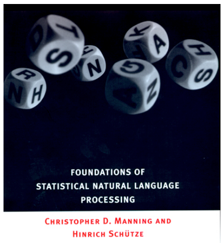
- Pros:
  - Pretty fast
  - ``Hot'' language for IR, NLP, ML, distributed computing, web development
  - Clean, transparent exception handling
  - Clean, minimalist syntax
  - Consistent semantics
  - Statically typed
  - Garbage collection
  - At least one great IDE (IntelliJ
  - Fully compatible with Java (use all Java libraries)
- Cons:
  - It has some "dark corners"
  - Backwards compatibility not guaranteed
  - No deep learning library native to Scala

Performance comparison



More benchmarks:
http://benchmarksgame.alioth.debian.org/u64/benchmark.php?test=all&lang=all&data=u64

Textbook



http://nlp.stanford.edu/fsnlp

I will provide all the other additional materials.

Grading

Component	Weight
Assignments	300 pts
Midterm exam	200 pts
Final exam	275 pts
Programming project	200 pts
In-class participation	25 pts
Total	**1000 pts**

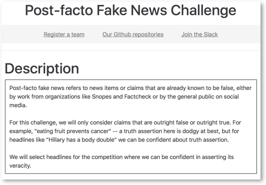
Grade	Point Range
A	900 – 1000
B	800 – 899
C	700 – 799
D	600 – 699
E	0 – 599

Four homeworks

Task	Deadline
HW 1	August 27
HW 2	September 24
Midterm review	October 10
Midterm	October 12
HW 3	October 29
HW 4	November 26
Final review	December 5
Project	December 7

Final project

- Implement a complete solution of a relevant NLP application or component.
- You can choose your own, but each must be validated by the instructor.
- For example:



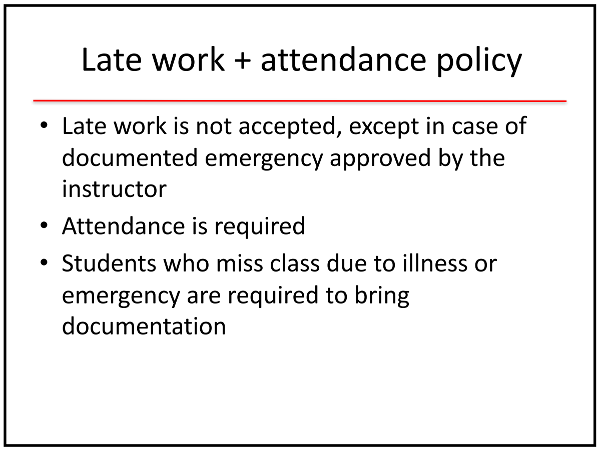
Post-facto Fake News Challenge

Description

Post-facto fake news refers to news items or claims that are already known to be false, either by work from organizations like Snopes and Factcheck or by the general public on social media.

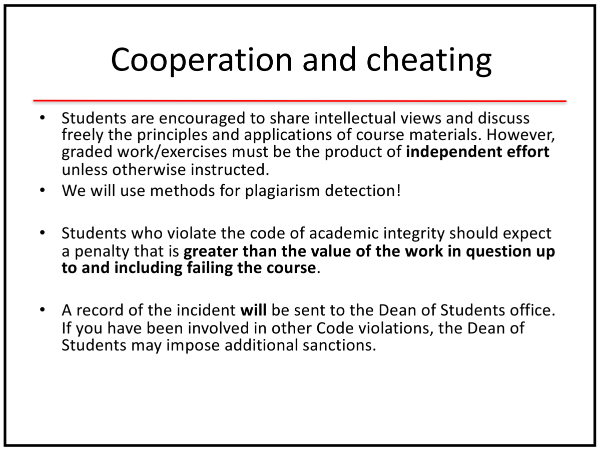
For this challenge, we will only consider claims that are outright false or outright true. For example, "eating fruit prevents cancer" -- a truth assertion here is dodgy at best, but for headlines like "Hillary has a body double" we can be confident about truth assertion.

We will select headlines for the competition where we can be confident in asserting its veracity.



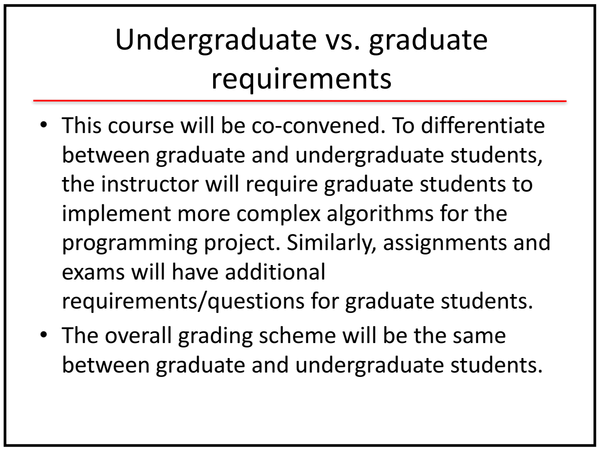
Late work + attendance policy

- Late work is not accepted, except in case of documented emergency approved by the instructor
- Attendance is required
- Students who miss class due to illness or emergency are required to bring documentation



Cooperation and cheating

- Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work/exercises must be the product of **independent effort** unless otherwise instructed.
- We will use methods for plagiarism detection!
- Students who violate the code of academic integrity should expect a penalty that is **greater than the value of the work in question up to and including failing the course**.
- A record of the incident **will** be sent to the Dean of Students office. If you have been involved in other Code violations, the Dean of Students may impose additional sanctions.



Undergraduate vs. graduate requirements

- This course will be co-convened. To differentiate between graduate and undergraduate students, the instructor will require graduate students to implement more complex algorithms for the programming project. Similarly, assignments and exams will have additional requirements/questions for graduate students.
- The overall grading scheme will be the same between graduate and undergraduate students.

Overview

- Administration
- First homework
- Course overview

THE GREATEST INSPIRATION IS THE DEADLINE

First homework

- **Due Sunday night (8/27)!**

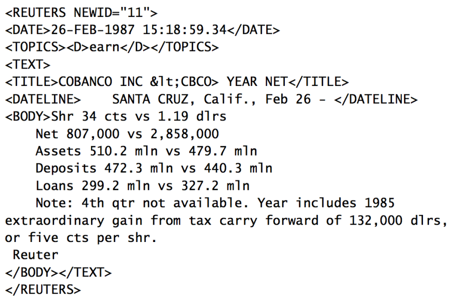
- Let’s take a look

Overview

- Administration
- First homework
- Course overview

PART 1: TEXT CATEGORIZATION AND A CRASH COURSE IN MACHINE LEARNING

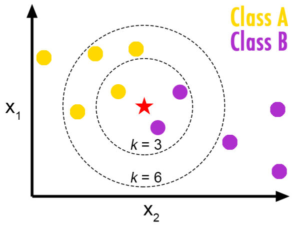
Text categorization



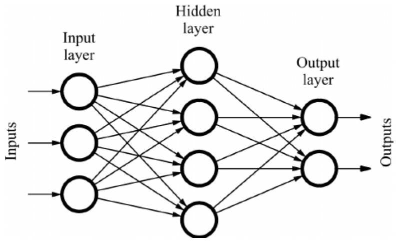
```
<REUTERS NEWID="11">
<DATE>26-FEB-1987 15:18:59.34</DATE>
<TOPICS><D>earn</D></TOPICS>
<TEXT>
<TITLE>COBANCO INC &lt;CBCO> YEAR NET</TITLE>
<DATELINE>    SANTA CRUZ, Calif., Feb 26 - </DATELINE>
<BODY>Shr 34 cts vs 1.19 dlrs
    Net 807,000 vs 2,858,000
    Assets 510.2 mln vs 479.7 mln
    Deposits 472.3 mln vs 440.3 mln
    Loans 299.2 mln vs 327.2 mln
    Note: 4th qtr not available. Year includes 1985
extraordinary gain from tax carry forward of 132,000 dlrs,
or five cts per shr.
 Reuter
</BODY></TEXT>
</REUTERS>
```

Other examples of text categorization?

Algorithms for classification: from kNN to feed-forward neural networks

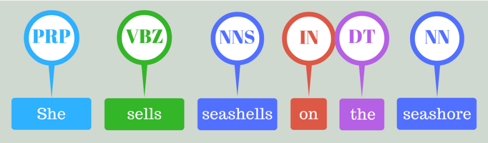


Algorithms for classification: from kNN to feed-forward neural networks



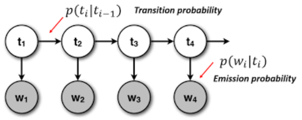
PART 2: SEQUENCE MODELS

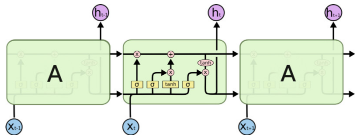
Part-of-speech tagging



Other examples of applications of sequence models?

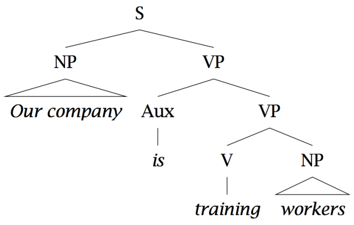
From hidden Markov models to long short-term memory models



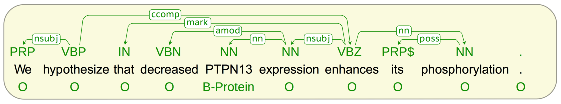


PART 3: PARSING

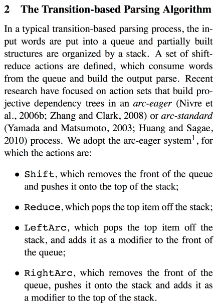
Constituent parsing



Dependency parsing


ccomp
mark
amod
nn
nsubj
nsubj
nn
poss
PRP VBP IN VBN NN NN VBZ PRP$ NN .
We hypothesize that decreased PTPN13 expression enhances its phosphorylation .
O O O O B-Protein O O O O O

Shift-reduce parsing



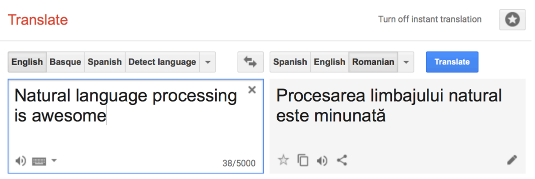
2 The Transition-based Parsing Algorithm

In a typical transition-based parsing process, the input words are put into a queue and partially built structures are organized by a stack. A set of shift-reduce actions are defined, which consume words from the queue and build the output parse. Recent research have focused on action sets that build projective dependency trees in an *arc-eager* (Nivre et al., 2006b; Zhang and Clark, 2008) or *arc-standard* (Yamada and Matsumoto, 2003; Huang and Sagae, 2010) process. We adopt the arc-eager system[1], for which the actions are:

- `Shift`, which removes the front of the queue and pushes it onto the top of the stack;
- `Reduce`, which pops the top item off the stack;
- `LeftArc`, which pops the top item off the stack, and adds it as a modifier to the front of the queue;
- `RightArc`, which removes the front of the queue, pushes it onto the stack and adds it as a modifier to the top of the stack.

PART 4: ALIGNMENT AND MACHINE TRANSLATION

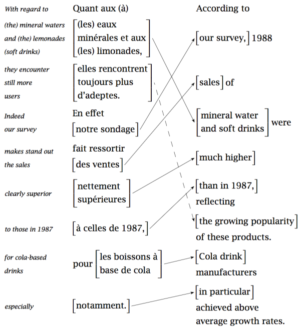
Machine translation



Natural language processing is awesome

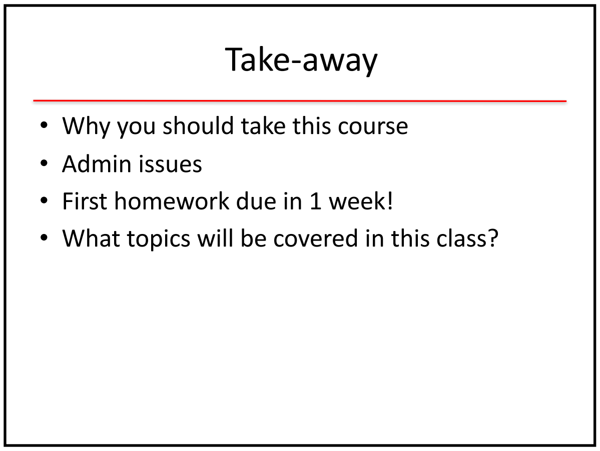
Procesarea limbajului natural este minunată

Alignment models



Part 5 (time permitting): Advanced techniques

PART 5 (TIME PERMITTING): ADVANCED TECHNIQUES



Take-away

- Why you should take this course
- Admin issues
- First homework due in 1 week!
- What topics will be covered in this class?