# Building a State-of-the-Art Grammatical Error Correction System

**Mr. Leuchtag** "Liebchen--sweetnessheart, what watch?"
**Mrs. Leuchtag** "Ten watch."
**Mr. Leuchtag** "Such much?"

-- *Casablanca*

-What time is it?
-Ten o'clock.
-So late?

Alla Rozovskaya Columbia University
Dan Roth University of Illinois

**NAACL**, 06/03/2015

# English today

Most of the text today is written in **English**


- As many as **2 billion speakers**
  - Billions of tweets and emails
  - Millions of scientific articles
- Over 75% of those writers are **non-native speakers** (Crystal'05)

# English as a Second Language (ESL) learners

- Existing spell-checkers cannot deal with mistakes typical for non-native speakers of English


- **Common mistake types:** *articles, prepositions, noun number, verb errors*

```
To my opinion, phone has many functions,

included camera and Wi-Fi receiver.
```

# Automated error correction

- Growing interest in the topic of error correction in the NLP community
  - **The need for text correction** in many areas

**Writing assistance software**

**Second language learning**

**Robust NLP tools for “noisy” data**

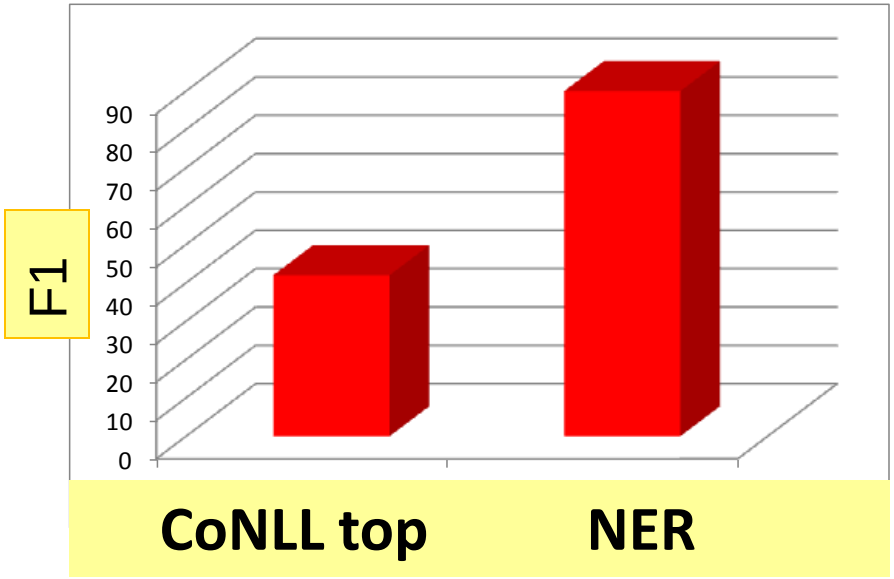
# Four text correction competitions in the NLP community

- **HOO competitions** (Dale and Kilgarriff'11, Dale et al.'12)
- **CoNLL shared tasks**
  - **CoNLL-2013** (Ng et al.)
    - **Articles, prepositions, noun number, 2 kinds of verb errors**
  - **CoNLL-2014** (Ng et al.)
    - All types of mistakes

# ESL error correction is a difficult task

**vs. other NLP tasks, e.g.**

**N**amed **E**ntity **R**ecognition




- **Performance of ESL writers seems high;** over 90% of words are used correctly
- Because of that, it is hard to improve over the learner texts

# Machine Translation

**Russian:** Грибов в лесу полным-полно.

**English:** There’re lots of mushrooms in the forest.

**English (Google Translate):** Mushrooms in a forest full of them.

# In this talk

- We show that the ESL error correction task can be successfully addressed through a combination of **machine learning techniques** and **linguistic knowledge**

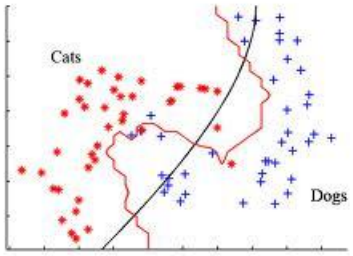
# Contributions

Analysis of the top system in the CoNLL-2013 competition

Key system dimensions

# Key system dimensions

Algorithmic perspective:
*ACL'11*




Choice of the training data

Model adaptation to learner errors:
*NAACL'10, ACL'11, BEA'12*

Linguistic knowledge

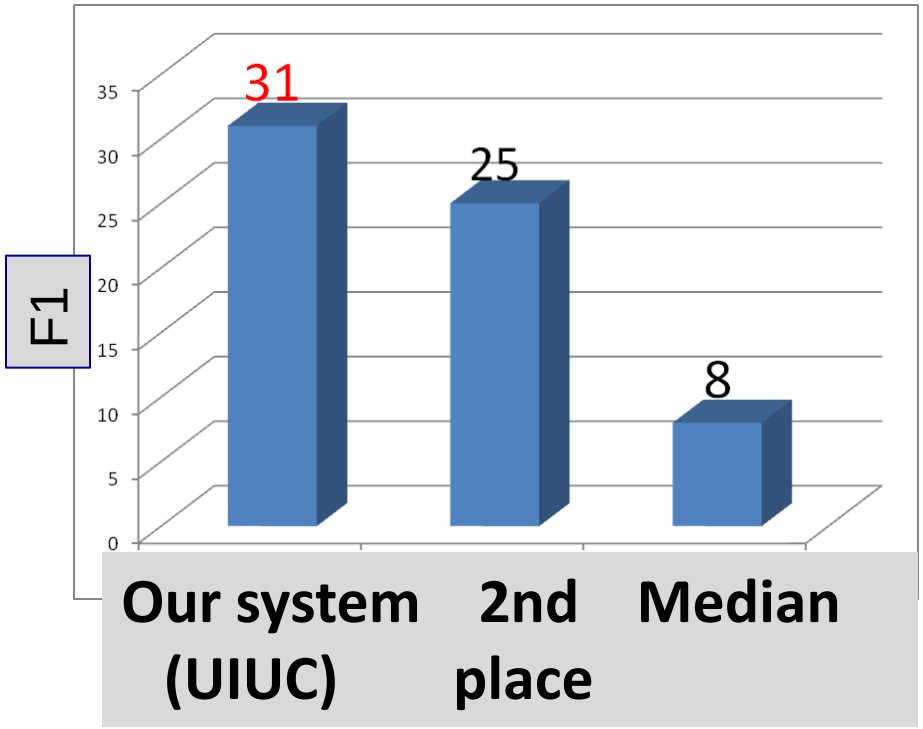
# Outline

- CoNLL-2013 overview
  - Top systems and approaches
  - Illinois system
  - System analysis
  - Conclusions

# CoNLL-2013 competition

- Data:
  - Essays written by ESL learners at NUS
    - **Training (learner) data:** 1.2 M words annotated for errors
- Focuses on 5 common error types:
  - Article, preposition, noun number, verb agreement, verb form
- Participants were allowed to use additional resources

# CoNLL-2013 shared task (top results)




**The 17 teams that participated used a variety of rule-based and statistical methods.**

# Top approaches

| System | Approach | Resources |
|---|---|---|
| **UIUC** | Machine-learning classifiers | Learner data, Web1T corpus |
| **NTHU** | Count-based model | Web1T corpus |
| **HIT** | Classifiers and rule-based methods | WordNet |
| **NARA** | SMT, classifiers, LMs | Gigaword, learner data |
| **UMC** | Rules, classifiers, LMs | Learner, Web1T |

Teams used very similar resources!

# Outline

- CoNLL-2013 overview
- Top systems and approaches
- Illinois system
- Analysis
- Conclusions

# The Illinois system

- 5 error-specific classifiers:

| Error | Machine learning approach | Training data | Adaptation | Ling. knowledge |
|---|---|---|---|---|
| **Article** | Discriminative (Averaged Perceptron) | Learner data | Yes, artificial errors | Features |
| **Preposition** | Naïve Bayes | Native (Web 1T) | Yes, priors | - |
| **Noun number, verb agreement, verb form** | Naïve Bayes | Native (Web 1T) | No | Candidate generation, verb finiteness |

# Outline

- CoNLL-2013 overview
- Top systems and approaches
- The  Illinois system
- Analysis
- Conclusions

# Dim 1: Choice of the learning algorithm (ACL'11)

- **Discriminative** model - **Averaged Perceptron (AP)**
- **Generative** model - **Naïve Bayes (NB)**
- **Language model (LM)**
  - Interpolated count-based LM with Jelinek-Mercer linear interpolation

# Key findings on the algorithm comparison

- The discriminative model is the best-performing model

**Averaged Perceptron (AP) > Naïve Bayes (NB) > Language Model (LM)**

- AP ≈ 2NB
- AP ≈ 5LM

*But sometimes we do not want to train discriminatively!*

# Choice of the learning algorithm

| Error | Classifier | F1 |
|---|---|---|
| **Article** | LM | 21.11 |
| | NB | **32.35** |
| **Preposition** | LM | 12.09 |
| | NB | **14.04** |
| **Noun number** | LM | 40.72 |
| | NB | **42.60** |
| **Verb agreement** | LM | 20.65 |
| | NB | **26.46** |
| **Verb form** | LM | 13.40 |
| | NB | **14.50** |

More experiments and results on another learner data set in the paper!

Training on native data (Web 1T corpus)

# Dim. 2: Native vs. learner data for training

**Trade-off**

- Size
- Type of information
- Different phenomena are in different sides of this tradeoff

# Dim. 2: Two types of information

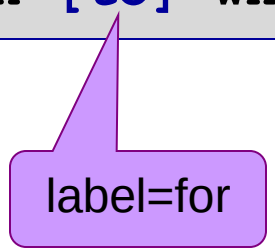
**He is an engineer with a passion [to] what he does.**

## (1) Context information

**He is an engineer with a passion [to] what he does.**

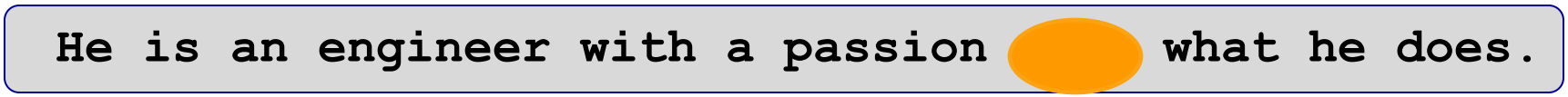
## (2) Author's word (which could be an error)

**He is an engineer with a passion [to] what he does.**



# Training on native data

- Decision is made only based on **context information**




- **Author's word** is not taken into consideration

# Training on learner data

- **Training** on learner data with annotated errors

**He is an engineer with a passion to what he does.**

**Context information**

label=for

**Author's word**

# Native vs. learner data for training

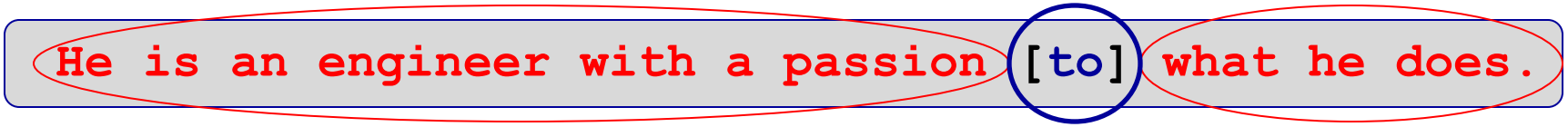
| Error | Train. data | F1 (%) |
|---|---|---|
| **Article** | Native | 34.49 |
| | Learner | 33.50 |
| **Preposition** | Native | 12.09 |
| | Learner | 10.26 |
| **Noun number** | Native | **42.60** |
| | Learner | 19.22 |
| **Verb agreement** | Native | 23.46 |
| | Learner | **27.93** |
| **Verb form** | Native | **18.35** |
| | Learner | 12.32 |

Some types of mistakes due to their nature require more data and thus especially benefit, when a model is trained on native data

# Dim. 3: Adaptation to learner errors

- Adaptation allows us to exploit advantages of the two training sources in one model!

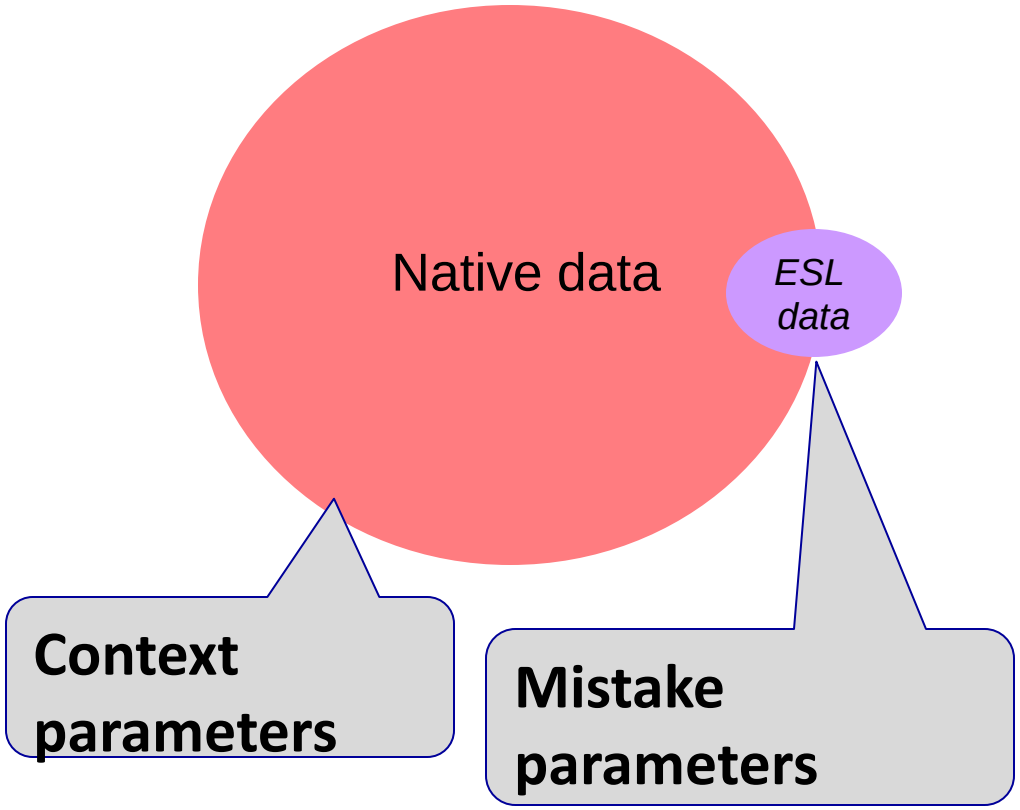
# Adaptation




- Context parameters are complex, so we need a lot of data for estimation
- But mistake parameters are simple!

# ESL adaptation

- A way of combining lots of native data with small amounts of annotated ESL data in training

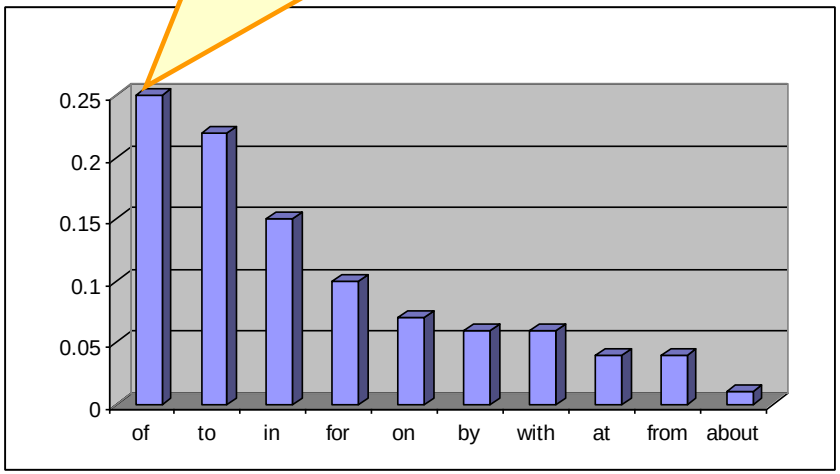


# Native-data priors vs. adapted priors

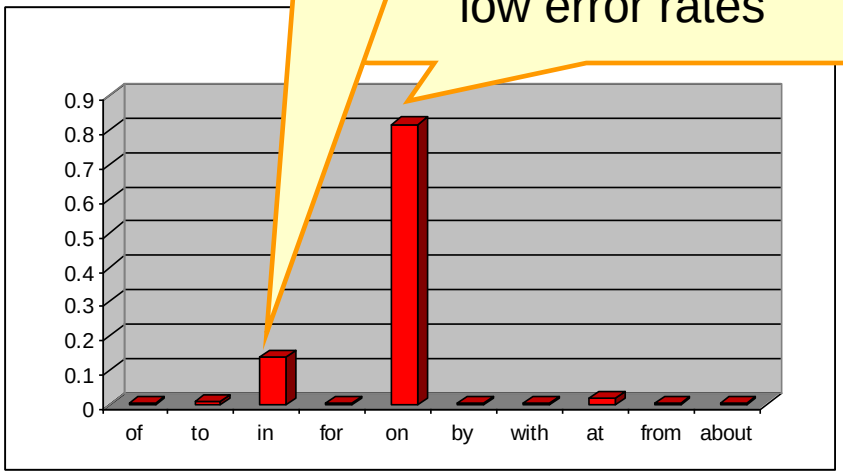
Highest prior for the most frequent preposition in the native data

Adapted priors reflect likely preposition confusions

Highest prior for the source; reflects the low error rates




**Priors based on native data**




**Adapted priors for the source "on"**

Native priors reflect preposition frequencies in native data; adapted priors reflect error rates and likely confusions.

# Adaptation with a small amount of annotation

Two methods for the top-performing models:

- **Artificial errors method** (for discriminative classifiers, NAACL'10, BEA'12)
- **Priors method** (for Naïve Bayes, ACL'11)

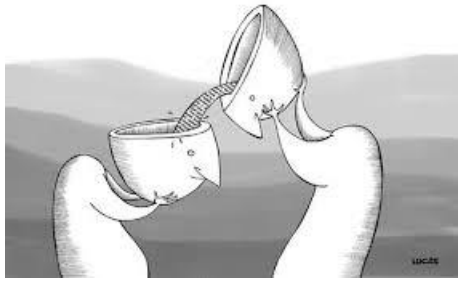
# NB adaptation

- Adaptation is useful for all errors, except for noun number

| Error | Algorithm | F1 (%) |
|---|---|---|
| **Article** | NB | 18.28 |
| | NB-adapted | **19.18** |
| **Preposition** | NB | 09.03 |
| | NB-adapted | **10.94** |
| **Noun number** | NB | **23.06** |
| | NB-adapted | 22.89 |
| **Verb agreement** | NB | 16.72 |
| | NB-adapted | **17.62** |
| **Verb form** | NB | 11.93 |
| | NB-adapted | **14.63** |

**NB adaptation**: results on the CoNLL training data.

# Dim 4: Linguistic knowledge



- Features (article and verb agreement)
- Candidate identification (nouns, verbs)
- Finiteness (verb errors, see EACL'14)

# Verb finiteness

Finite

We **discuss** this every time.

Non-finite

They let us **discuss** this.

Grammatical properties associated with each type are mutually exclusive

# Using verb finiteness to correct verb errors (EACL'14)

| Training method | F1 (%) |
|---|---|
| One classifier | 16.43 |
| Separate finiteness-based training | **21.08** |

**Verb agreement and verb form errors:** Improvement due to separate training

# Conclusion

- ESL error correction is an important problem
  - Many applications: e.g. educational technology, data analytics
- The approach I presented is based on:
  - Understanding the linguistic aspects of the task
  - Matching them with the appropriate machine learning solutions

Thank you!