# Natural Errrror Processing

**Mariana Romanyshyn**

Technical Lead, Sr. Computational Linguist
Grammarly, Inc.

# Contents

# 1. Define the problem

I like
cooking my family
and my pets.

Use commas.
Don't be a psycho.

WRITING.COM
WHERE THE WRITERS GO

# Some statistics

Non-natives make **1** mistake every **10** words.

Most popular errors:

- spelling
- preposition choice
- missing article
- missing or redundant punctuation
- word choice

Irony
is when
someone writes
“Your an idiot”.

Learn grammar.
Insult properly.

WRITING.COM
WHERE THE WRITERS GO

# Tasks:

- Detection
- Correction

# Tasks:

- Detection
- Correction

He is just want us to make a good decision.
Today my plan is study the system.
John is now be able to donate money.
It is say that their house is sold.

is + VB => ?

~~is just want~~ → just wants

~~is study~~ → is to study

~~is now be~~ → is now

~~is say~~ → is said


ATTENTION
Toilet
ONLY
for
DISABLED
ELDERLY
PREGNANT
CHILDREN

# 2. Classical methods: rules and ngrams

# Basic Text Pre-processing

- Language identification
- Paragraph splitting
- Sentence splitting
- Tokenization
- POS tagging

# 2.1. Tokens and parts of speech

- **MD ? RB {of->have} VBN**
  - “could {of->have} done”
  - “could n’t {of->have} done”
- **CD “[ap]\\.?m\\.?” ~~(time-expression)~~**
  - 10 am ~~in the morning~~
  - 7:30 p.m. ~~in the evening~~
- **“the” {“most” JJS->JJS|most JJ}**
  - the {most nicest->nicest}
  - the {most beautifullest->beautiful}

# 2.2. Dictionaries

- **(Dr|Mr|Mrs|...) |.| (not |.|)**
  - “Dr->Dr. Stevenson”
- **over-regularized verb: get infinitive and transform it properly**
  - “I {eated->ate} your cookie”
  - “I have {eated->eaten} your cookie”
- **slang**
  - “Do you {wanna->want to} watch TV?”
  - “I really {wanna->want} this dress.”
  - “I really {wanna->want an} apple.”

# 2.3. Ngrams (1)

## Double article

- **Easy case:**
  - {An a->An} apple a day keeps the doctor away.
  - Please book {a the->the} following flights for me.
- **Multiple choice:**
  - Give me {the a->the/a} chance to solve the problem.
  - They all have access to {a the->the} internet.
  - I slipped {the a->a} couple magazines under the bed.
- **Misspellings:**
  - Jane, these {a the->are the} 3 pictures.
  - {A the->At the} beginning, we were together.
  - Hunter is a freshmen {an a->and a} basketball player.

# 2.3. Ngrams (2)

- **Jane, these {a the->are the} 3 pictures.**
  - ngram(“these and the 3 pictures”) = 100
  - ngram(“these are the 3 pictures”) = **5,000**
  - ngram(“these as the 3 pictures”) = 50
  - ngram(“these at the 3 pictures”) = 30

- **{A the->At the} beginning, we were together.**
  - ngram(“<S> And the beginning ,”) = 100
  - ngram(“<S> Are the beginning ,”) = 0
  - ngram(“<S> As the beginning ,”) = 60
  - ngram(“<S> At the beginning ,”) = **80,000**

# 2.4. Syntactic trees (1)

**Barry**, the guy I met yesterday, who has three kids, **{live->lives}** in Brooklyn.

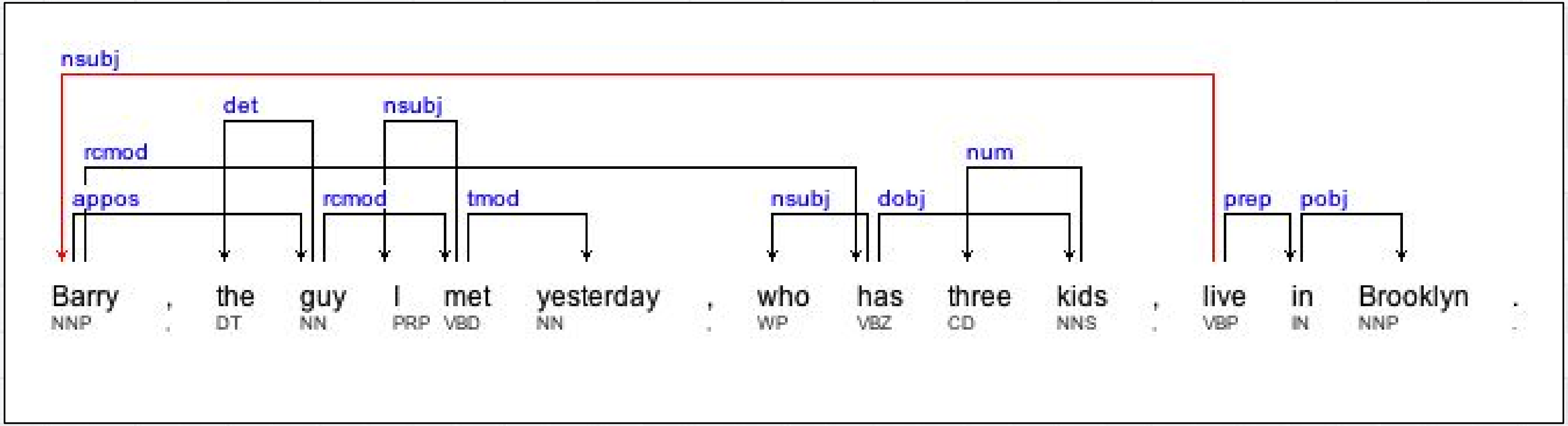
# 2.4. Syntactic trees (2)

**Barry**, the guy I met yesterday, who has three kids, **{live->lives}** in Brooklyn.

- Too far for POS tags
  - **NNP** , DT NN PRP VBD , WP VBZ CD NNS , **VBP**
- Too far for ngrams
  - ngrams("kids , **live** in Brooklyn") = 0
  - ngrams("kids , **lives** in Brooklyn") = 0

# 2.4. Syntactic trees (3)

**Barry**, the guy I met yesterday, who has three kids, **{live->lives}** in Brooklyn.

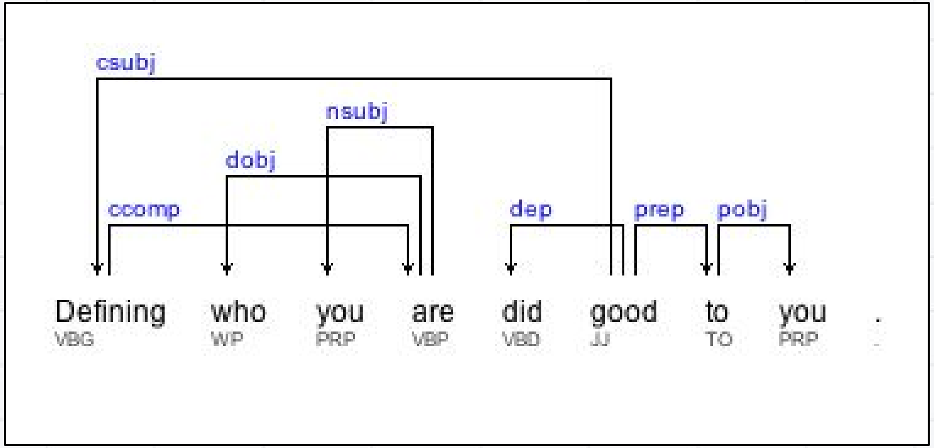


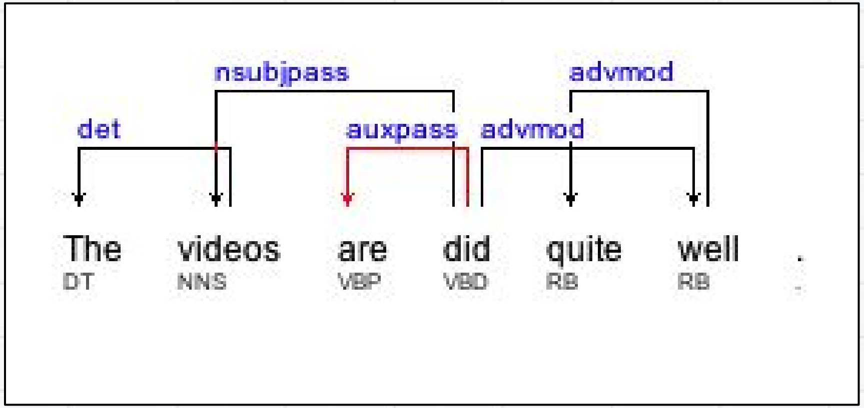
# 2.4. Syntactic trees (4)

+ Defining who you **are did** good to you.
- The videos **are did** quite well.

# 2.4. Syntactic trees (5)

- \+ Defining who you **are did** good to you.
- \- The videos **are did** quite well.



# Statistics repeated

Non-natives make **1** mistake every **10** words.

Most popular errors:

- spelling
- preposition choice
- missing article
- missing or redundant punctuation
- word choice

# 3. Machine Learning

# 3.1. What we need (1)

Two things needed:

1. A machine learning classifier (MaxEnt, SVM, Naive Bayes, Random Forest, Average Perceptron, etc.)

2. Data with labels for each training example

# 3.1. What we need (2)

Two things needed:

1. A machine learning classifier (MaxEnt, SVM, Naive Bayes, Random Forest, Average Perceptron, etc.)

2. Data with labels for each training example

**What features to use? ¯_(ツ)_/¯**

# 3.2. Preposition choice

# Preposition choice: tricky cases

- **Multiple corrections**
  - The economic globalization is of the most concern {by->to/of/for} each nation.
  - We have problems {such as/like/with} rapid development.

- **Correct but rare usage**
  - I rely mostly {upon->on} my instinctive feeling.

# Preposition choice: features

- **Lexical features**
  - word
  - left/right context
- **Grammatical features**
  - part of speech
  - dependency relations
  - constituency spans
- **Ngrams**
  - unigrams, bigrams, three-grams…

# Preposition choice: features

- **More ngrams**
  - POS ngrams
    - ngrams(“decided_VBD **for_IN** going_VBG to_IN”)
    - ngrams(“decided_VBD **on_IN** going_VBG to_IN”)
  - Wildcard ngrams
    - ngrams(“decided for **VBG** to”)
    - ngrams(“decided on **VBG** to”)

# Preposition choice: features

- **More ngrams**
  - POS ngrams
    - ngrams(“decided_VBD **for_IN** going_VBG to_IN”)
    - ngrams(“decided_VBD **on_IN** going_VBG to_IN”)
  - Wildcard ngrams
    - ngrams(“decided for **VBG** to”)
    - ngrams(“decided on **VBG** to”)

# Preposition choice: features

- **Semantic features**
  - WordNet
  - VerbNet
  - semantic role labelling

# Preposition choice: features

- **Semantic features**
  - WordNet
  - VerbNet
  - semantic role labelling
- **Linguistic resources**
  - governing dictionaries
  - word-form dictionaries

# Preposition choice: features

- **Semantic features**
  - WordNet
  - VerbNet
  - semantic role labelling
- **Linguistic resources**
  - governing dictionaries
  - word-form dictionaries
- **Sources**
  - L1 of the writer
  - genre of the writing

# 3.3. Overly complex words

# Complex words: tricky cases

- **Complex word in complex sentences**
  - The researchers used optical coherence tomography to elucidate the impact of fixation on retinal laser pathology.

- **Compound words**
  - Snow, raindrops, hail, or sleet that fall from above can be collected in hydrosphere alleys.

# Complex words: features

- **Lexical features**
    - word
    - left/right context
- **Grammatical features**
    - part of speech
    - dependency relations
    - constituency spans
- **Ngrams**
    - unigrams, bigrams, three-grams...

# Complex words: features

- **More Ngrams**
  - character ngrams
    - procrastinate: *procr, rocra, ocras, crast, rasti, astin…*

# Complex words: features

- **More Ngrams**
  - character ngrams
    - procrastinate: *procr, rocra, ocras, crast, rasti, astin…*
- **Spelling of the word**
  - Is the word capitalized?
  - Is the word hyphenated?

# Complex words: features

- **More Ngrams**
  - character ngrams
    - procrastinate: *procr, rocra, ocras, crast, rasti, astin…*
- **Spelling of the word**
  - Is the word capitalized?
  - Is the word hyphenated?
- **Morphology**
  - Is the word compound?
  - Which affixes are common for complex words?

# Complex words: features

- **Number of senses**

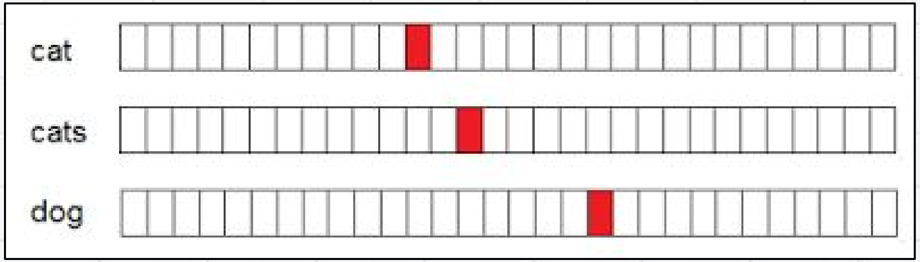
| Word | Number of senses in WordNet |
|---|---|
| report | 7 n + 6 v |
| cat | 8 n + 2 v |
| elucidate | 2 v |
| procrastinate | 2 v |
| moribund | 2 a |

# Complex words: features

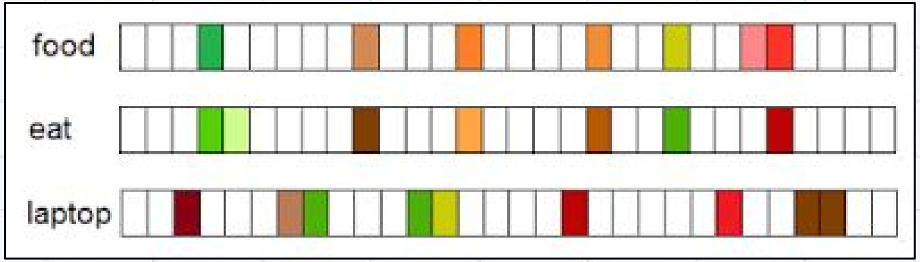
- **More Features**
  - word length
  - ratio of vowels vs consonants
  - number of syllables
  - word position in the sentence
  - depth of the word in the dependency tree of the sentence
  - degree of concreteness/abstractness using MRC Psycholinguistic Database...

# 3.4. Word representations (1)

- One-hot vectors: the word in the vocabulary



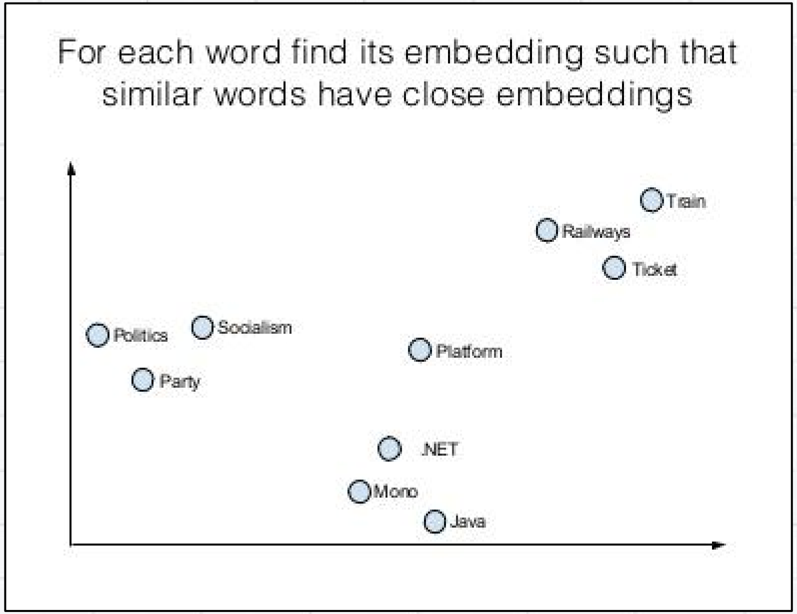

- Distributional vectors: the context of the word



# 3.4. Word representations (2)

- Word embeddings
  - start from a random vector for each word
  - observe the target word and its contexts
  - for each context, update the word's vector *

*...so that the contexts predict the word better



# 4. Is there more?

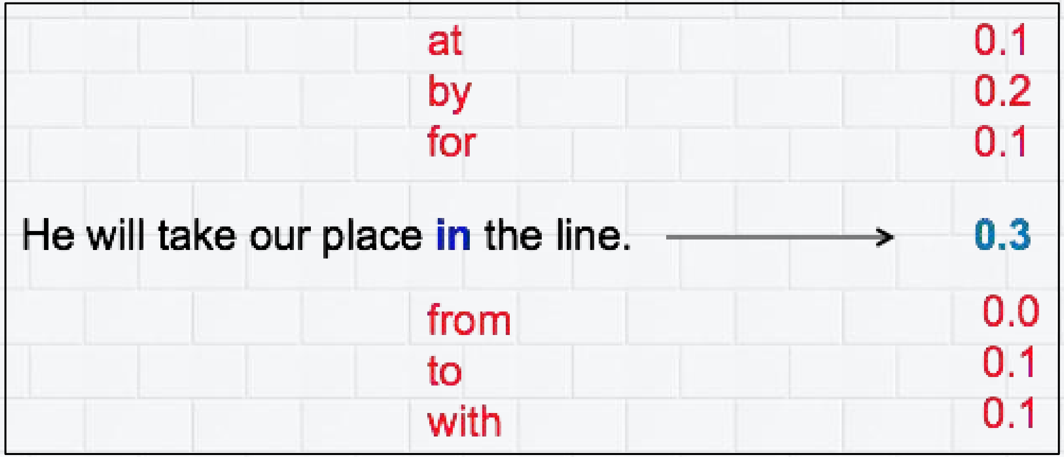
# Yes!

# 4.1. Language Modelling

Shows the probability of a sentence or phrase in the language.
Usually trained on good genre-balanced texts.

Used:

- for error detection
- for error correction
- as a feature for ML
- as a filter



# 4.2. Grammaticality Classifier

Shows the probability of an error in the sentence.
Trained on error-labelled sentences.

Used for:

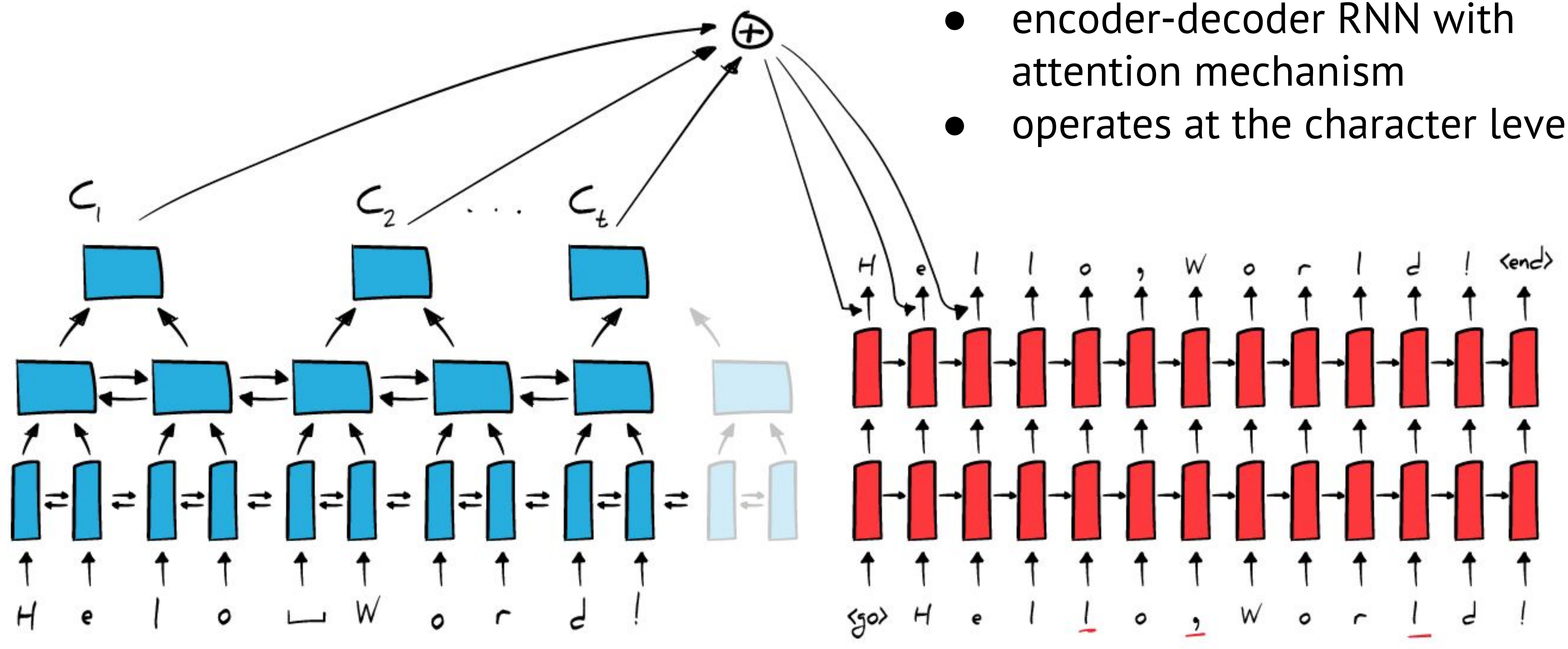
- selecting sentences for error correction
- filtering out false positives

# 4.3. Machine Translation

Two types:

- Noisy channel
  - translation from “bad English” to “good English”
- Round-trip
  - translation from English to another language and then back to English

# 4.4. Neural Error Correction (2016)




- encoder-decoder RNN with attention mechanism
- operates at the character level

# 5. Where do I get data?


The data says we need more data.
someecards
user card

# 5.1. Free Data

- **Open Data**
  - Wikipedia
  - Wiktionary
  - Common Crawl
  - Specific language resources
- **Artificial Data**
  - take “correct” corpora
  - plant errors

# 5.2. Paid Data

- **Licensed data**
  - for error correction: NUCLE, Lang-8, CLC, CLEC, ICLE
- **Annotated data**
  - pricey: Appen
  - cheaper: Amazon Mechanical Turk, CrowdFlower


Woo-Hoo!
Data!

# 6. How to evaluate the solution?

# 6.1. Traditional Formulas

$$Precision = \frac{TPs}{TPs + FPs}$$

$$F\text{ - }score = \frac{Precision * Recall}{Precision + Recall}$$

$$Recall = \frac{TPs}{TPs + FNs}$$

$$Accuracy = \frac{TPs + TNs}{TPs + TNs + FPs + FNs}$$

# 6.2. Preposition correction (2014)

| Team | Description | Recall, % |
|---|---|---|
| CAMB | Rule-based → LM ranking → SMT → LM ranking | **38.63** |
| NTHU | Language modelling. | 20.42 |
| AMU | Phrase-based MT with augmented LMs and specific features. | 18.41 |
| CUUI | A combination of averaged perceptron, naive Bayes, and pattern-based learning. | 18.22 |
| UMC | Factored MT model using modified POS tags and morphology as features. | 16.98 |
| SJTU | Rule-based error categorization used to train a Maximum Entropy model. | 8.95 |
| POST | Detection via n-grams; correction via LM alternatives. Post-processing with rules. | 2.25 |

# 6.3. Quality on NUCLE (2014)

| Team | Description | F-score, % |
|---|---|---|
| CAMB | Rule-based → LM ranking → SMT → LM ranking | 37.33 |
| CUUI | A combination of averaged perceptron, naive Bayes, and pattern-based learning. Check-specific features. | 36.79 |
| AMU | Phrase-based MT with augmented LMs and specific features. | 35.01 |
| POST | Detection via n-grams; correction via LM alternatives. Post-processing with rules. | 30.88 |
| NTHU | Language modelling, conditional random field model, rules, MT for different checks. | 29.92 |
| UMC | Factored MT model using modified POS tags and morphology as features. | 25.37 |
| SJTU | Rule-based error categorization used to train a Maximum Entropy model. | 15.19 |

# 6.3. Quality on NUCLE (2014)

| Team | Description | F-score, % |
|---|---|---|
| **NEC** | **Attention-based encoder-decoder RNN and a 5-gram LM. (2016)** | 40.56 |
| CAMB | Rule-based → LM ranking → SMT → LM ranking | 37.33 |
| CUUI | A combination of averaged perceptron, naive Bayes, and pattern-based learning. Check-specific features. | 36.79 |
| AMU | Phrase-based MT with augmented LMs and specific features. | 35.01 |
| POST | Detection via n-grams; correction via LM alternatives. Post-processing with rules. | 30.88 |
| NTHU | Language modelling, conditional random field model, rules, MT for different checks. | 29.92 |
| UMC | Factored MT model using modified POS tags and morphology as features. | 25.37 |
| SJTU | Rule-based error categorization used to train a Maximum Entropy model. | 15.19 |

# Any kwestions?


Mariana Romanyshyn
mariana.romanyshyn@grammarly.com
mariana.scorp@gmail.com