# Automatic Grammatical Error Correction using Machine Translation

Roman Grundkiewicz
romang@inf.ed.ac.uk

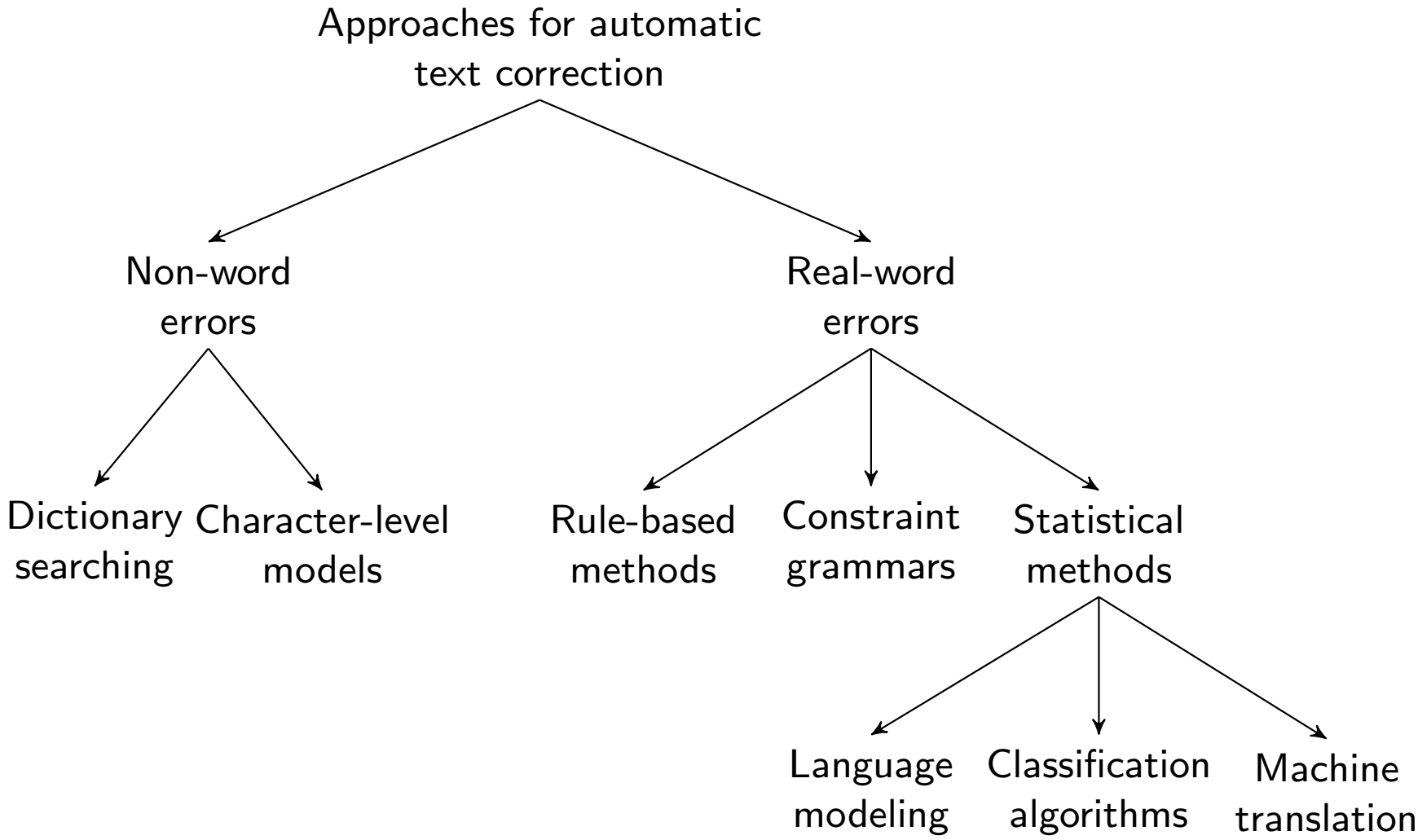
Adam Mickiewicz University in Poznan, University of Edinburgh

January 29, 2018

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645487 (MMT) and No 644333 (TraMOOC).
Project partially funded by the National Science Centre, Poland (Grant No. 2014/15/N/ST6/02330).

## History and today

- Unix Writer's Workbench, IBM's CRITIQUE
- Aspell, Ispell, ...


Approaches for automatic text correction
Non-word errors
Real-word errors
Dictionary searching
Character-level models
Rule-based methods
Constraint grammars
Statistical methods
Language modeling
Classification algorithms
Machine translation

## Shared tasks on GEC

- Helping Our Own (2011, 2012)
- CoNLL shared task: GEC (2013, 2014)
- QALB task on automatic text correction of Arabic (2014, 2015)
- Automated Evaluation of Scientific Writing (2016)

## CoNLL 2014 shared task: Grammatical Error Correction

Goal: correct errors in short English texts written by non-native speakers of English

- The NUS Corpus of Learner English (NUCLE)
- The MaxMatch ($M^2$) metric with $F_{0.5}$

## The NUS Corpus of Learner English (NUCLE)

- 1,414 essays (57,151 sentences) written by Singaporean ESL learners
- Grammatical errors manually corrected by professional English teachers
- 27 predefined error types

## NUCLE example

```
S When we are diagonosed out with certain genetic
  disease , are we suppose to disclose this result
  to our relatives ?

A 3 4|||Mec|||diagnosed|||REQUIRED|||-NONE-|||0
A 4 5|||Prep||||||REQUIRED|||-NONE-|||0
A 8 9|||Nn|||diseases|||REQUIRED|||-NONE-|||0

A 3 5|||Rloc-|||diagnosed|||REQUIRED|||-NONE-|||1
A 6 6|||ArtOrDet|||a|||REQUIRED|||-NONE-|||1
A 12 13|||Vt|||supposed|||REQUIRED|||-NONE-|||1
```

We focus on errors produced by English as a Second Language (ESL) learners.

The modern GEC task aims to correct all types of errors: grammatical, stylistic, spelling, etc.

Why is it difficult?

## The MaxMatch ($M^2$) metric

$M^2$ computes the sequence of phrase-level edits between writer's sentence and a system output ($e_i$) that achives the highest overlap with the gold standard annotation ($g_i$).

*There is no* **a doubt**, *tracking* **system has** *brought many benefits in this information age.*

$g_i = \{\text{a doubt} \rightarrow \text{doubt}\}, \{\text{system} \rightarrow \text{systems}\}, \{\text{has} \rightarrow \text{have}\}$
$e_i = \{\text{a} \rightarrow \emptyset\}$ $e_i = \{\text{a doubt} \rightarrow \text{doubt}\}$

$$P = 1/1 \quad R = 1/3 \quad F_{0.5} = 5/7$$

| Rank | Team ID | P | R | $M^2_{0.5}$ |
|---|---|---|---|---|
| 1 | CAMB* | 39.71 | 30.10 | 37.33 |
| 2 | CUUI | 41.78 | 24.88 | 36.79 |
| 3 | AMU** | 41.62 | 21.40 | 35.01 |
| 4 | POST | 34.51 | 21.73 | 30.88 |
| 5 | NTHU* | 35.08 | 18.85 | 29.92 |
| 6 | RAC | 33.14 | 14.99 | 26.68 |
| 7 | UMC** | 31.27 | 14.46 | 25.37 |
| 8 | PKU | 32.21 | 13.65 | 25.32 |
| 9 | NARA | 21.57 | 29.38 | 22.78 |
| 10 | SJTU | 30.11 | 5.10 | 15.19 |
| 11 | UFC | 70.00 | 1.72 | 7.84 |
| 12 | IPN | 11.28 | 2.85 | 7.09 |
| 13 | IITB* | 30.77 | 1.39 | 5.90 |

The official CoNLL-2014 shared task results

Why phrase-based SMT?

- No restriction to specific error types
- No annotations of error types required
- Correction of errors within phrases
- Correction of interacting errors within a sentence
- Easy incorporation of monolingual data
- No expert linguistic knowledge is required
- Relatively easy adaptation to other languages

Grammatical Error Correction using

# Statistical Machine Translation

Grammatical error correction as a kind of translation from "incorrect" English to "correct" English.

## Phrase-based statistical machine translation

$$\hat{T} = \arg\max_{T} p(T|S)$$

- Sequences of words (phrases) as atomic units
- Phrase translations and their probabilities are learnt from parallel data

Log-linear model:

$$\log p(T|S) = \sum_{i=1}^{N} \lambda_i \log h_i(T|S)$$

Common features:

- $h_{\mathrm{PT}}(T|S) \approx \sum_a \prod_{j=1}^{|T|} p(s_{(a)j}|t_i)$
- $h_{\mathrm{LM}}(T) \approx \prod_{i=1}^{|T|} p(t_i|t_{i-n+1}, \ldots, t_{i-1})$

Weight parameters $\lambda_i$ tuned on held-out data according to the choosen evaluation metric.

## Task-specific features

1. Stateless dense features — the word-level Levenshtein distance and edit operation counts between phrases
2. Stateful dense features — $n$-gram language models on different factors
3. Sparse features — correction patterns on words with context

Edit-based features

| Source phrase | Target phrase | $d_{LD}$ | $d_D$ | $d_I$ | $d_S$ |
|---|---|---|---|---|---|
| a short time . | short term only . | 3 | 1 | 1 | 1 |
| a situation | into a situation | 1 | 0 | 1 | 0 |
| a supermarket . | a supermarket . | 0 | 0 | 0 | 0 |
| a supermarket . | at a supermarket | 2 | 1 | 1 | 0 |

$$\log p(T|S) = \sum_{i=1}^{N} \lambda_i \log h_i(T|S) + \lambda_D h_D(T|S) + \lambda_I h_I(T|S) + \lambda_S h_S(T|S).$$

$N$-gram Language Model (LM):

- $p_{\text{LM}}(T) \approx \prod_{i=1}^{|T|} p(t_i | t_{i-n+1}, \ldots, t_{i-1})$

Word Class Language Model (WCLM):

- `word2vec`, 200 classes:
  $C_1 = \{is, was, has, had, about, became, did, began, \ldots\}$
  $C_{36} = \{of, for, with, on, from, at, after, into, \ldots\}$

Operation Sequence Model (OSM):

- *"people are more health conscious"*
  $\rightarrow$ *"people have been more health conscious than before"*:

  `_COPY_`people `_TRANS_`are`_TO_`have `_INS_`been `_COPY_`more
  `_COPY_`health `_COPY_`conscious `_INS_`than `_INS_`before

## Experiment settings

- Moses SMT toolkit [Koehn et al., 2007]
- Word-alignments produced with MGIZA++ [Gao and Vogel, 2008]
- Weight optimization with Minimum Error Rate Training (MERT) [Och, 2003]
- Training data: NUCLE + Lang-8

Eleison Oct 1, 2013 20:46

I was sufferring from a nightmare last night.

I was suffering from a nightmare last night.

This works but just "I had a nightmare last night" might sound more natural

1 people think this correction is good.

I was standing on a long line.

I was standing in a long line.

1 people think this correction is good.

Demons are cuttting off a head of people in line in order.

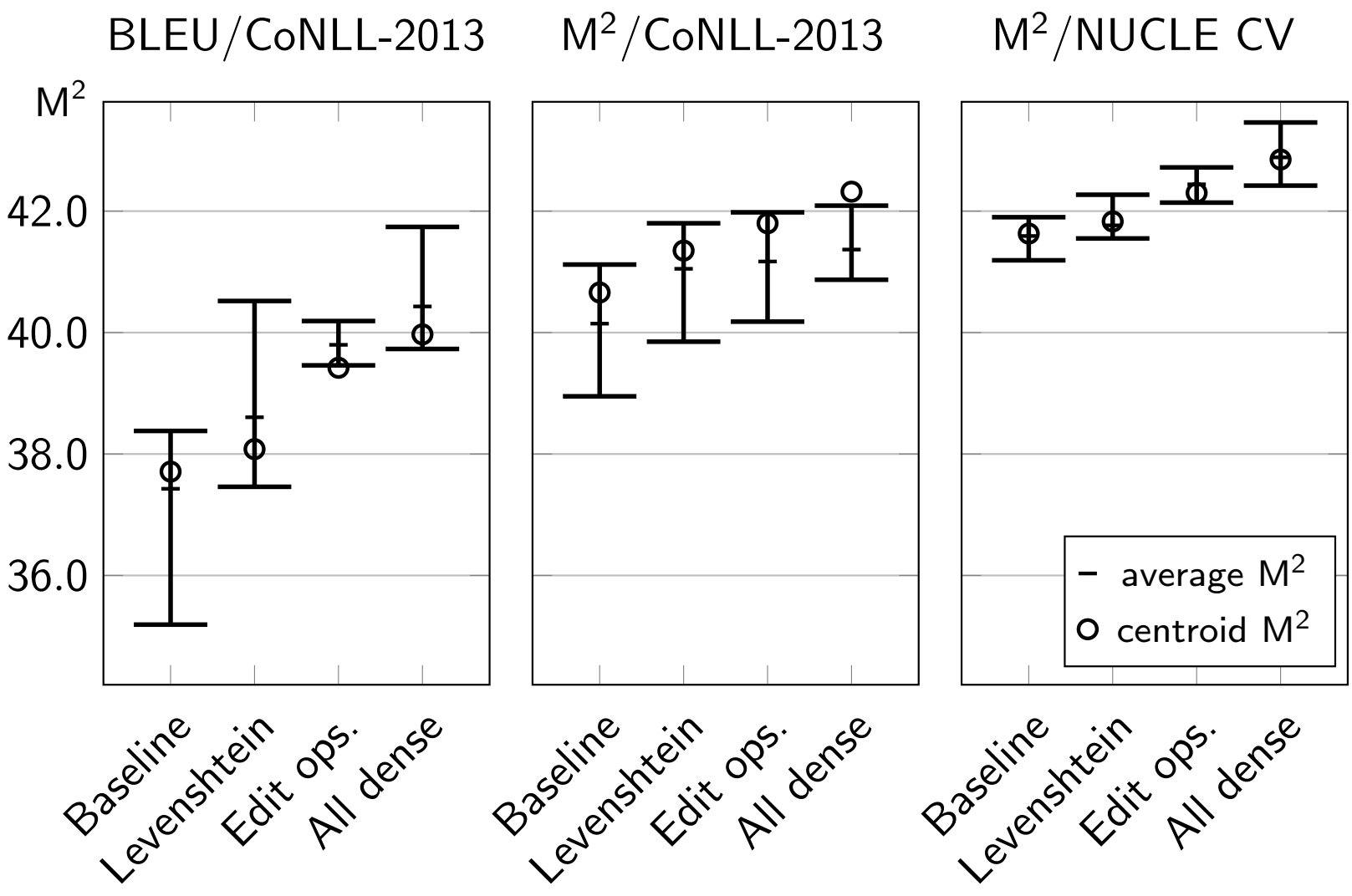
Demons are cutting off the heads of people in line in order.

I was seeing heads roling down and waiting for my turn.

I saw heads rolling down and was waiting for my turn.

Parallel and monolingual data

| Corpus | Sentences | Tokens |
|---|---|---|
| NUCLE | 57,151 | 1,161,567 |
| CoNLL-2013 Test Set | 1,381 | 29,207 |
| CoNLL-2014 Test Set | 1,312 | 30,144 |
| Lang-8 NAIST | 2,186,460 | 25,732,858 |
| Wikipedia | 213.08 M | 3.37 G |
| CommonCrawl (u) | 59.13 G | 975.63 G |


BLEU/CoNLL-2013
M2/CoNLL-2013
M2/NUCLE CV
M2
42.0
40.0
38.0
36.0
average M2
centroid M2
Baseline
Levenshtein
Edit ops.
All dense

## NUCLE as a development set

- 4-fold cross validation
  - 3 parts added to the training data, 1 part used as a tune set
  - Increased error rate in the tune set to ca. 15%
  - Training 4 models, re-tuning each model 5 times
- Training the final model on all training data
- Averaging weights vectors

## Sparse features

A large number of binary-valued features to make the model aware of specific phenomena:

$$h_{CP,j}(T|S) = \begin{cases} 1 & \text{if } c_j \in C(T,S) \\ 0 & \text{otherwise} \end{cases}.$$

## Sparse features (E0)

*"Then a new problem comes out ."*
$\rightarrow$ *"Hence , a new problem surfaces ."*

```
sub(Then,Hence)
ins(,)
sub(comes,surfaces)
del(out)
```

## Sparse features (E0C10)

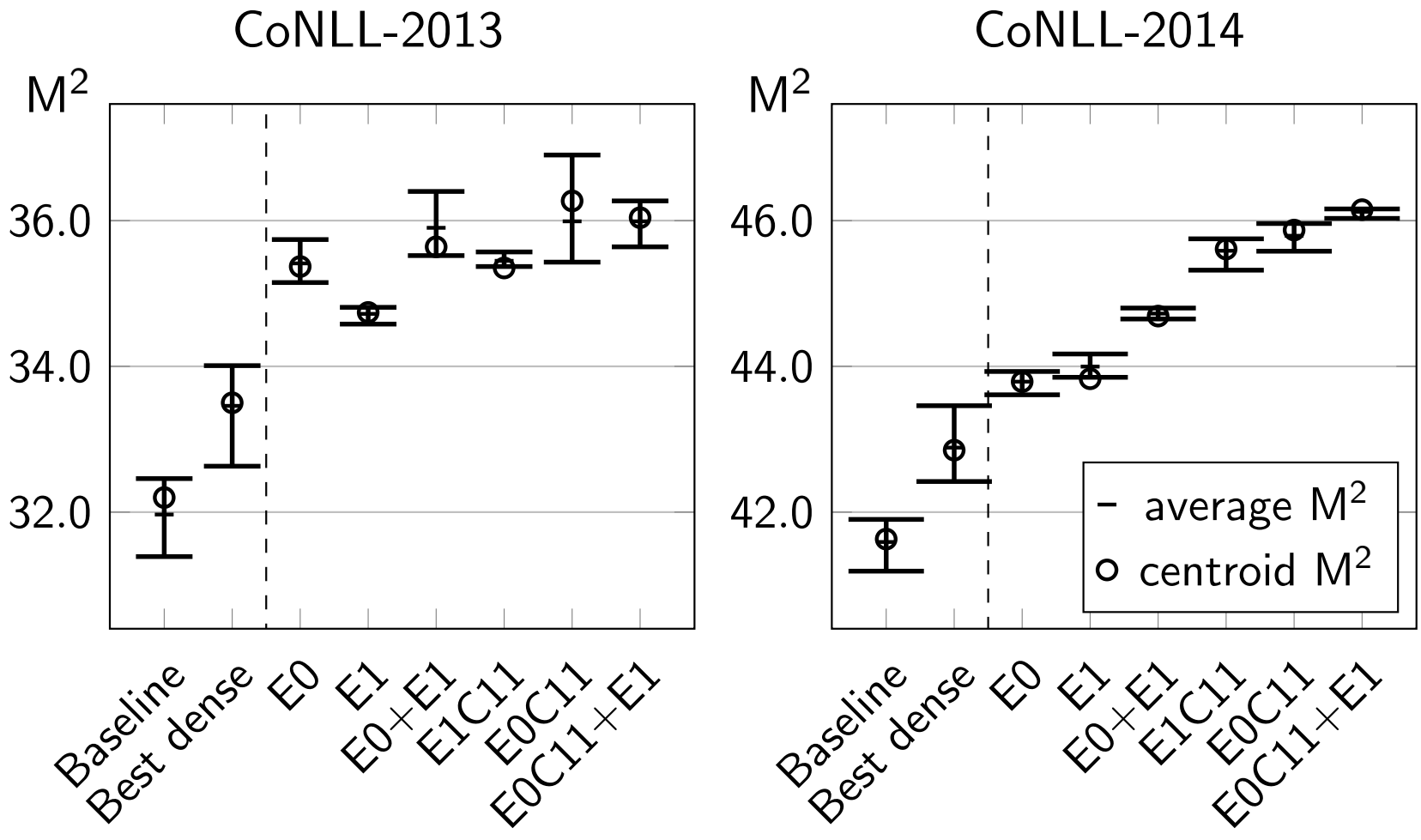
*"Then a new problem comes out ."*
$\rightarrow$ *"Hence , a new problem surfaces ."*

```
<s>_sub(Then,Hence)
sub(Then,Hence)_a
Hence_ins(,)
ins(,)_a
problem_sub(comes,surfaces)
sub(comes,surfaces)_out
comes_del(out)
del(out)_.
<s>_sub(Then,Hence)_a
Hence_ins(,)_a
problem_sub(comes,surfaces)_out
comes_del(out)_.
```

## Sparse features (E0C11)

*"Then a new problem comes out ."*
$\rightarrow$ *"Hence , a new problem surfaces ."*

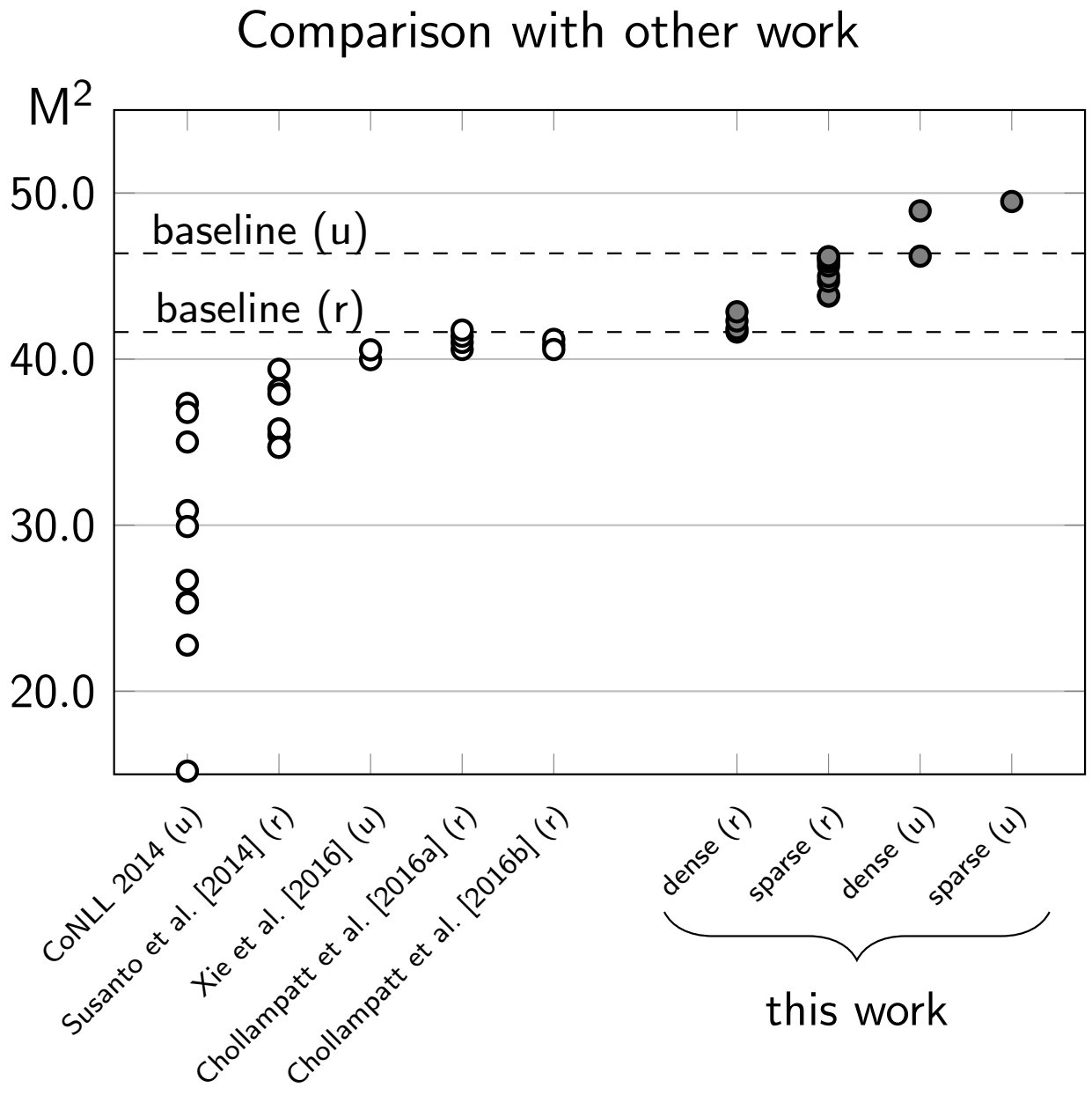
```
0_sub(Then,Hence)
sub(Then,Hence)_2
130_ins(,)
ins(,)_2
57_sub(comes,surfaces)
sub(comes,surfaces)_36
1_del(out)
del(out)_0
<s>_sub(Then,Hence)_2
130_ins(,)_2
57_sub(comes,surfaces)_36
1_del(out)_0
```




Results for SMT systems with sparse features

## Adding a large-scale language model

| System | CoNLL-2014 Prec. | CoNLL-2014 Recall | CoNLL-2014 $M^2$ | CoNLL-10 Prec. | CoNLL-10 Recall | CoNLL-10 $M^2$ |
|---|---|---|---|---|---|---|
| Baseline | 48.97 | 26.03 | 41.63 | 69.22 | 37.00 | 58.95 |
| +CCLM | **58.91** | 25.05 | 46.37 | 76.66 | 36.39 | 62.77 |
| Best dense | 50.94 | 26.21 | 42.85 | 71.11 | 37.44 | 60.27 |
| +CCLM | 59.98 | 28.17 | **48.93** | 79.76 | 39.52 | 66.27 |
| Best sparse | 57.99 | 25.11 | 45.95 | 76.48 | 35.99 | 62.43 |
| +CCLM | 61.27 | 27.98 | **49.49** | 80.57 | 39.74 | **66.83** |


Comparison with other work
M2
50.0
40.0
30.0
20.0
baseline (u)
baseline (r)
CoNLL 2014 (u)
Susanto et al. [2014] (r)
Xie et al. [2016] (u)
Chollampatt et al. [2016a] (r)
Chollampatt et al. [2016b] (r)
dense (r)
sparse (r)
dense (u)
sparse (u)
this work

Grammatical Error Correction

# Summary

## Summary

Adopted SMT methods were weakly explored:

- Unknown hyperparameter settings
- No proper baselines

Some consequences:

- Underestimated potential of phrase-based machine translation
- Overestimated potential of classifier-based approaches

## Conclusions

- Weight optimization is important!
- Take advantage of easily-available monolingual data
- Task-specific features can help
- Ensembling methods

## Future work

- Neural GEC methods
  Data augmentation, domain adaptation, handling rare words (spelling errors), task-specific features, reinforcement learning, etc.
- Evaluation of high-performance GEC systems

S. Chollampatt, D. T. Hoang, and H. T. Ng. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1911, Austin, Texas, November 2016a. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1195`.

S. Chollampatt, K. Taghipour, and H. T. Ng. Neural network translation models for grammatical error correction. *arXiv preprint arXiv:1606.00189*, 2016b.

Q. Gao and S. Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. ACL, 2008.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics, 2007.

F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, USA, 2003. Association for Computational Linguistics.

R. H. Susanto, P. Phandi, and H. T. Ng. System combination for grammatical error correction. pages 951–962, 2014. URL `http://emnlp2014.org/papers/pdf/EMNLP2014102.pdf`.

Z. Xie, A. Avati, N. Arivazhagan, D. Jurafsky, and A. Y. Ng. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*, 2016.