# Clinical Spelling Correction for English and Dutch with Word and Character n-gram Embeddings

{ Pieter Fivez, Simon Šuster, Stéphan Tulkens and Walter Daelemans } CLiPS, University of Antwerp

## Contribution

- ✓ **context-sensitive** spelling correction for non-word misspellings in clinical text
- ✓ uses **word and character n-gram embeddings** to estimate semantic fit
- ✓ **outperforms** a state-of-the-art noisy channel model for **English**
- ✓ **transfers** well to **Dutch**

## Motivation

**clinical text** is characterized by

- high spelling **error** rate (mostly non-word)
- **variable** lexical characteristics
- **limited** accessible data

Applying the noisy channel model directly to this domain raises **three issues**

1. **Handling context-specificity**
   *iron deficiency due to ~~enemia~~ → anemia*
   *fluid injected with ~~enemia~~ → enema*
2. **Handling pseudo-random typos**
   *patient not ~~awre~~ → are* (correct: *aware*)
3. **Handling imbalanced or highly variable training data**
   → skewed corpus frequencies of domain-specific terms

## Acknowledgements

This research was carried out in the framework of the Accumulate IWT SBO project, funded by the government agency for Innovation by Science and Technology (IWT). We also like to thank Elyne Scheurwegs for preparing the Dutch data.

## References

[1] Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
[2] Lai, K. H., M. Topaz, F. R. Goss, and L. Zhou (2015). Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics 55*, 188–195.

## Concept

Rank misspelling replacement candidates according to **semantic fit**

→ extract salient clues from **misspelling context**

*iron deficiency due to ~~enemia~~ → anemia*

*fluid injected with ~~enemia~~ → enema*

→ **cosine similarity** between misspelling context and candidate in embedding space

**Embedding model**: Facebook fastText [1]

→ Word representations composed from word vectors and character n-gram vectors

→ constructs vectors for **unobserved** word types
→ more **generalized** representations
→ incorporates distributional semantic information from **subword** level

## Development

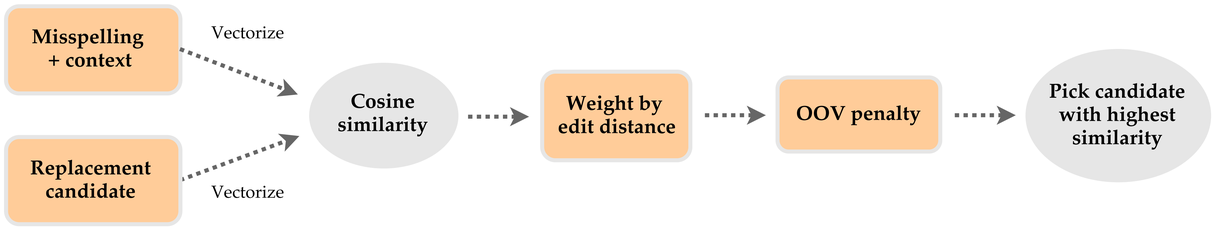
### Corpus creation

- 12,740 instances from the MIMIC-III corpus
- transforming randomly sampled in-vocabulary words with random Damerau-Levenshtein edit operations
- 80% 1 edit operation: *anemia → enemia*
- 20% 2 edit operations: *anemia → enemea*

### Experimental parameters

1. **Vector composition functions**
   → e.g. addition, multiplication
2. **Context weighting**
   → e.g. left vs. right window, window size
3. **Orthographic and phonetic edit distance**
4. **Out-of-vocabulary (OOV) penalty**
   → for candidate vectors constructed solely from character n-gram vectors

## Ranking Model Architecture




## Test Setting

### Models

**ClinSpell**: our context-sensitive clinical spelling correction method

**Noisy Channel**: our implementation of the state-of-the-art noisy channel method for clinical spelling correction by [2]

### Corpora

**English**: 555 annotated observed non-word misspellings from MIMIC-III

**Dutch**: 266 annotated observed non-word misspellings from heterogeneous clinical data from the University Hospital of Antwerp (UZA)

## Test Results

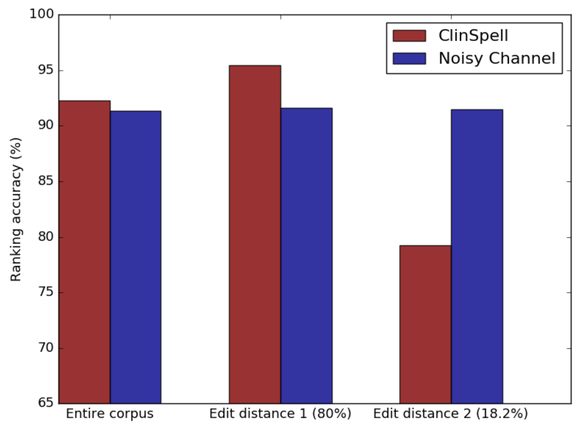



**Figure 1:** English ranking accuracies

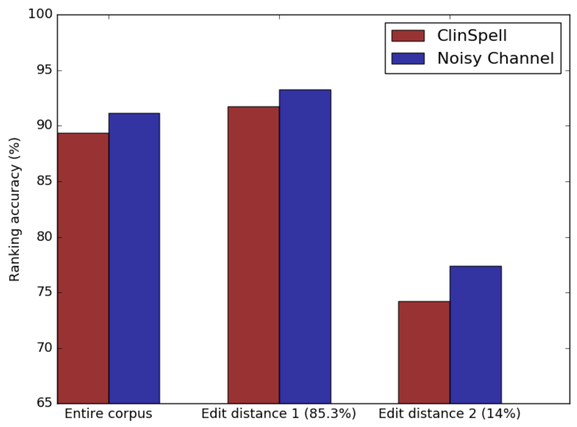



**Figure 2:** Dutch ranking accuracies

## Conclusions

+ Salient **performance gain** for English misspellings of 1 edit distance
+ **Contextual clues** can counter frequency bias
- Comparable performance when directly transferred to Dutch, but no superior performance yet → parameter tuning