1. Automatic classification of normalization replacement categories.
2. Distant supervision for normalization
3. Parsing social media data

Automatic classification of normalization replacement categories.

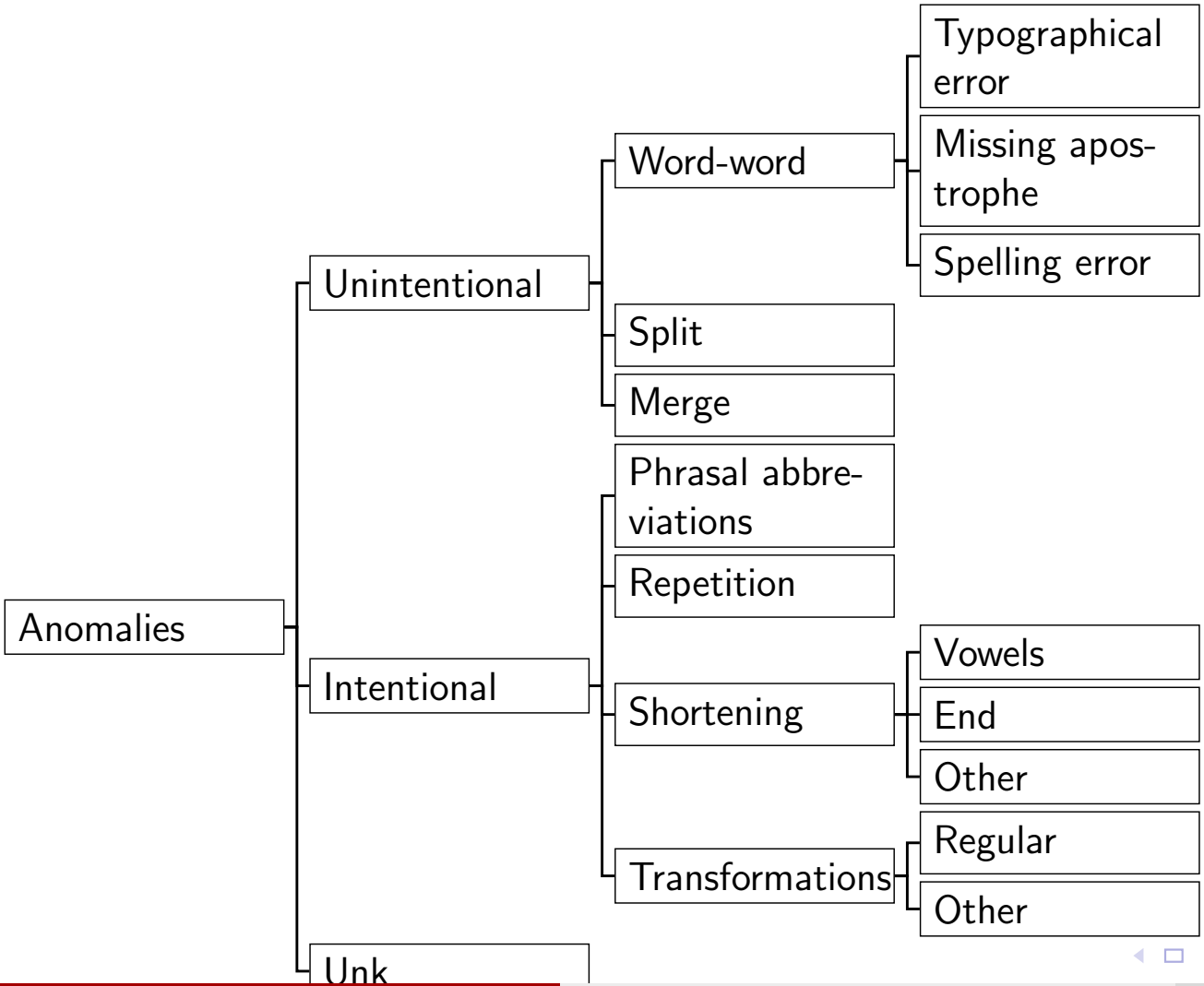
What?

orig	I	teering	up	from	alletgies	,	cant	tell	why
norm	i	tearing	up	from	allergies	,	can't	tell	why
ann		Spelling			Typo		Missin'		

orig	LMAO
norm	laughing my ass off
ann	Phrasal abbr.

Automatic classification of normalization replacement categories.

What?



Automatic classification of normalization replacement categories.

What?

- Three corpora in English, one already annotated

Automatic classification of normalization replacement categories.

What?

- Three corpora in English, one already annotated
- Corpora available in Dutch/German/Spanish/Slovenian/Serbian/Croation (Turkish-German code switched) ...

Automatic classification of normalization replacement categories.

Why?

- Compare corpora
- Evaluate normalization models in more detail for other languages as well
- Having this classification allows us to exclude categories from the model

Automatic classification of normalization replacement categories.

How?

- Classification problem with 14 classes
- Which classifier?
- Which features? (Across languages)

Automatic classification of normalization replacement categories.

More info:
Rob van der Goot, Rik van Noord and Gertjan van Noord. 2018. A Taxonomy for In-depth Evaluation of Normalization for User Generated Content. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation.

Christopher Bryant Mariano Felice Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Distant supervision for normalization

What?

- Automatically creating training data for a normalization system
- Existing normalization system can be used

Distant supervision for normalization

What? Find these automatically

orig	I	teering	up	from	alletgies	,	cant	tell	why	LMAO
norm	i	tearing	up	from	allergies	,	can't	tell	why	laughi

Distant supervision for normalization

Why?

- Allows for semi-supervised lexical normalization
  - Can normalize more languages
  - Can easily retrain for new domain/time-span

Distant supervision for normalization

How?

- For example by using brown clusters

Distant supervision for normalization

How?

- For example by using brown clusters
- Brown cluster for tomorrow:

2m, 2ma, 2mar, 2mara, 2maro, 2marrow, 2mor, 2mora, 2moro, 2morow, 2morr, 2morro, 2morrow, 2moz, 2mr, 2mro, 2mrrw, 2mrw, 2mw, tmmrw, tmo, tmoro, tmorrow, tmoz, tmr, tmro, tmrow, tmrrow, tmrrw, tmrw, tmrww, tmw, tomaro, tomarow, tomarro, tomarrow, tomm, tommarow, tommarrow, tommoro, tommorow, tommorrow, tommorw, tommrow, tomo, tomolo, tomoro, tomorow, tomorro, tomorrw, tomoz, tomrw, tomz

Distant supervision for normalization

Challenge: not all categories are captured by brown clusters.

Distant supervision for normalization

Use supplementary methods:
Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection. Sudhanshu Kasewa, Pontus Stenetorp and Sebastian Riedel. 2018. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Parsing social media data

- Popular task: last year three new treebanks for English
- Tokenization
- Differences in annotation
- ...