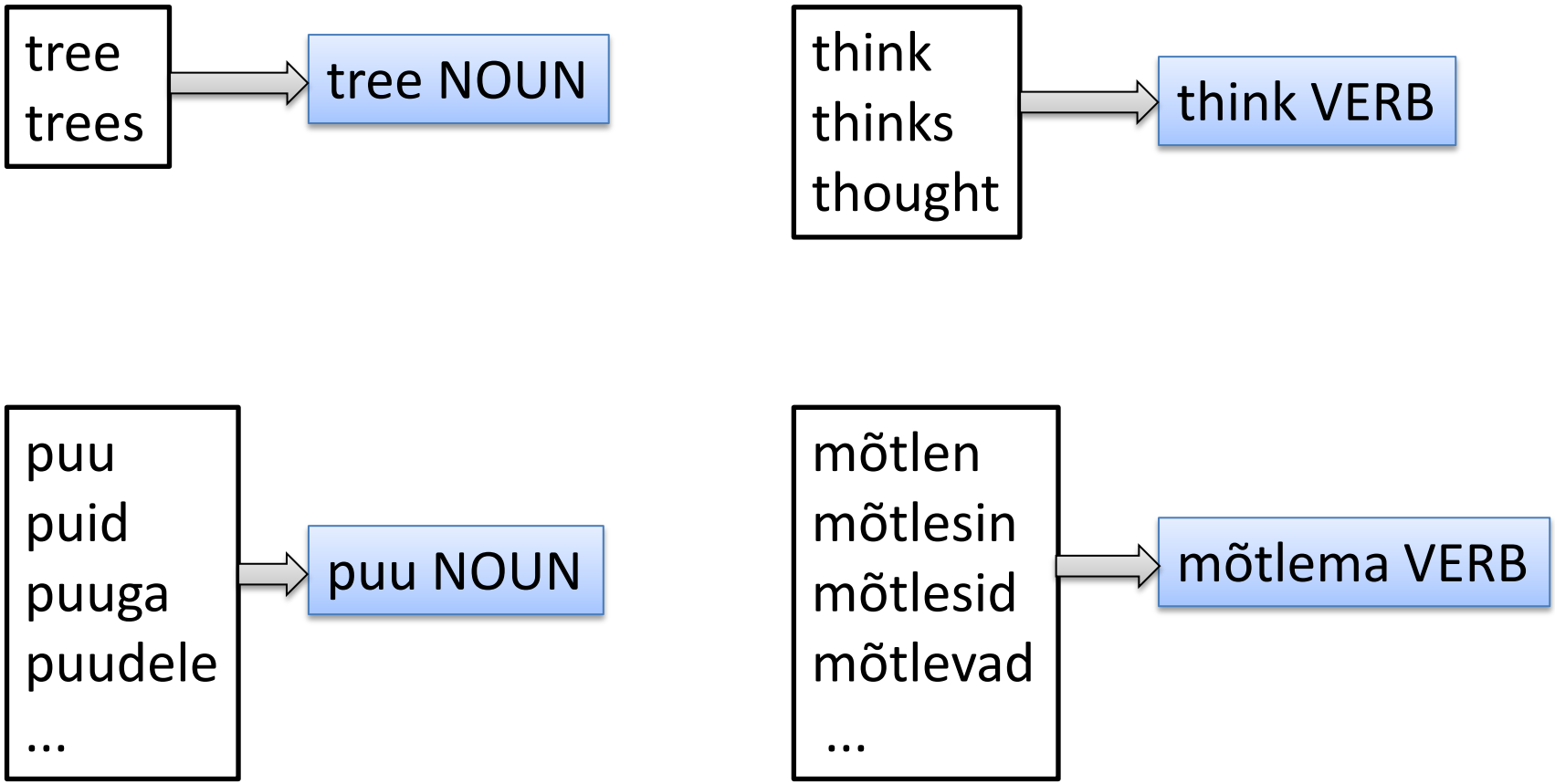
Noisy-Channel Spelling Correction Models for Estonian Learner Language Corpus Lemmatisation

Kairit Sirts
Institute of Cybernetics at TUT

Baltic HLT 2012
04.10.2012

Lemmatisation



Estonian Learner Language Corpus

- Created and developed in Tallinn University
- ca 500000 running words
- 93% first language Russian
- 66% students
- 44% essays, 18% questions-answers

Language learners make mistakes

Incorrect: mängisime erinevate pillede pael
Correct: mängisime erinevate pillide peal

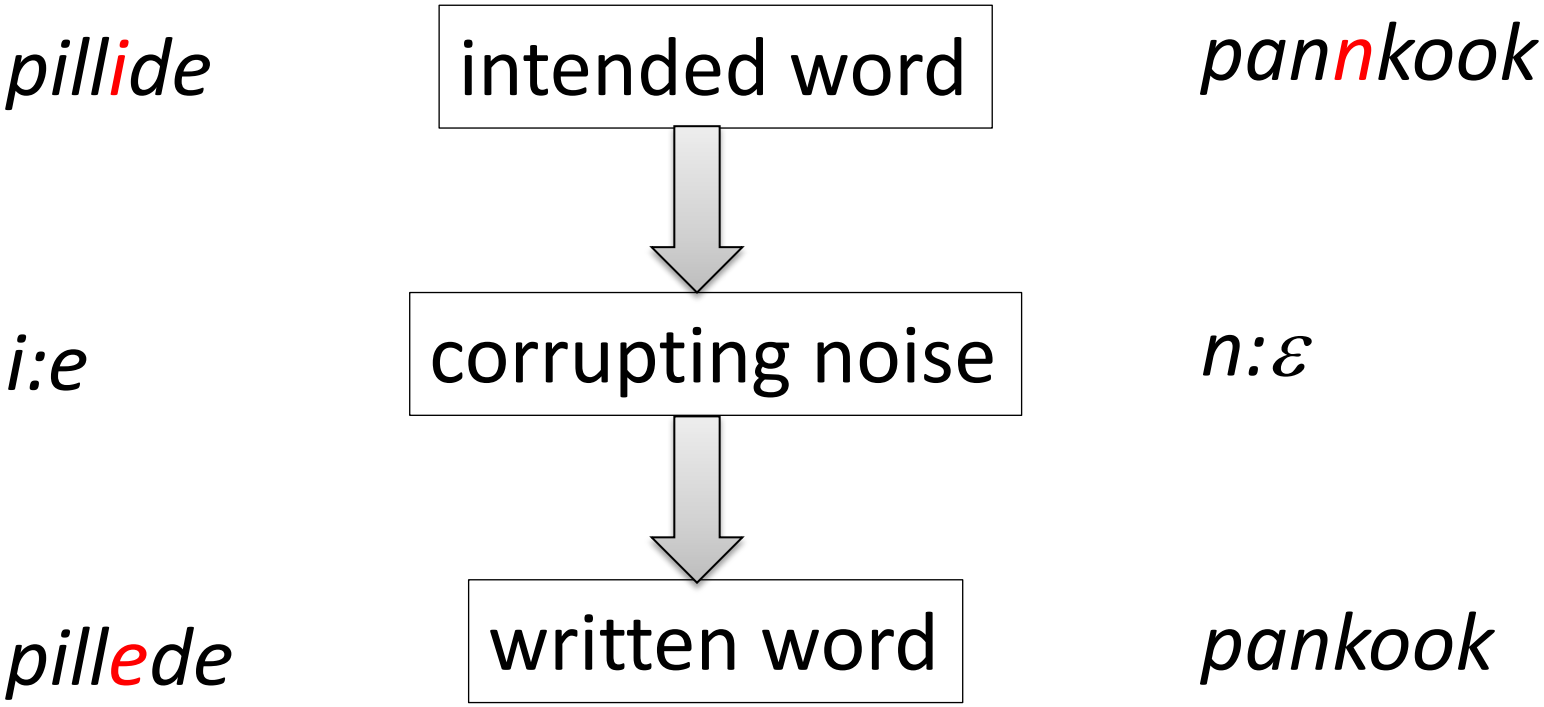
Incorrect: sõime seal suuri pankooke
Correct: sõime seal suuri pannkooke

Incorrect: me ei peame jalast käia
Correct: me ei pea jala käima

Learner Language Lemmatization

TOKEN	LEMMA	POS
mängisime	mängima	VERB
erinevate	erinev	ADJ
pillede	**Pille?**	**PROPER NAME?**
pael	**pael?**	**NOUN?**
sõime	sööma	VERB
seal	seal	ADVERB
suuri	suur	ADJ
pankooke	**pankook?**	**NOUN?**

Noisy-channel model



Noisy-channel Model

Language Model

$$\arg\max_{W} P(W \mid S) = \arg\max_{W} P(S \mid W) P(W)$$

Error Model

W – word
S – spelling

Correction-lemmatisation strategies

- Non-word spelling correction

  thes --> these, pankooke --> pannkooke

- Real-word spelling correction

  hole --> hope, pael --> peal

- Pipeline model
- Marginalized model
- Marginalized HMM

Correction-lemmatisation strategies

NON-WORD PIPELINE	NON-WORD MARGINALIZED	NON-WORD HMM
REAL-WORD PIPELINE	REAL-WORD MARGINALIZED	REAL-WORD HMM

Pipeline Model

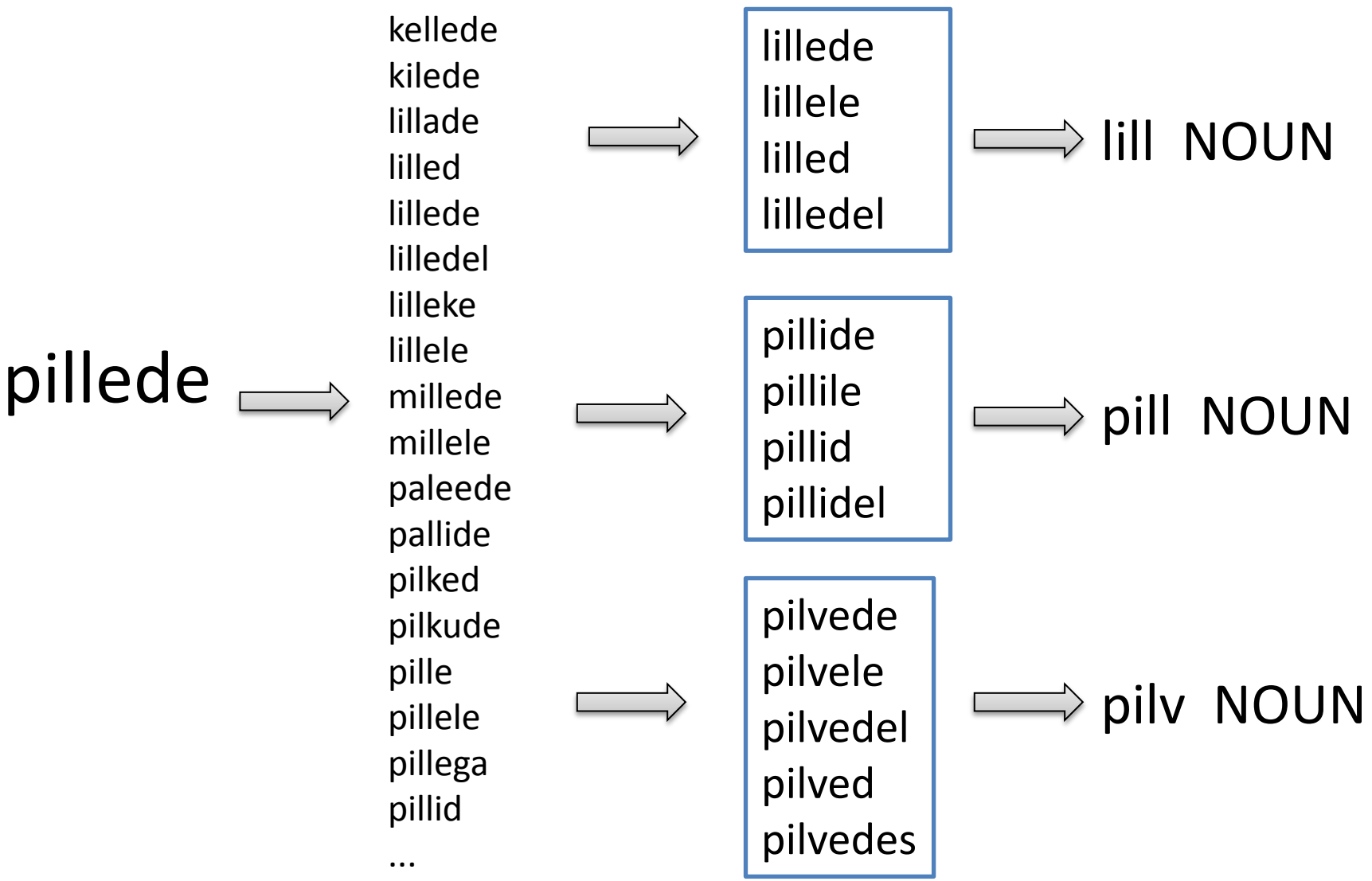
First spelling correction

Then lemmatisation with existing tools

$$\arg\max_{w} P(W \mid S) = \arg\max_{w} P(S \mid W) P(W)$$

W – word
S – spelling

Marginalized Model



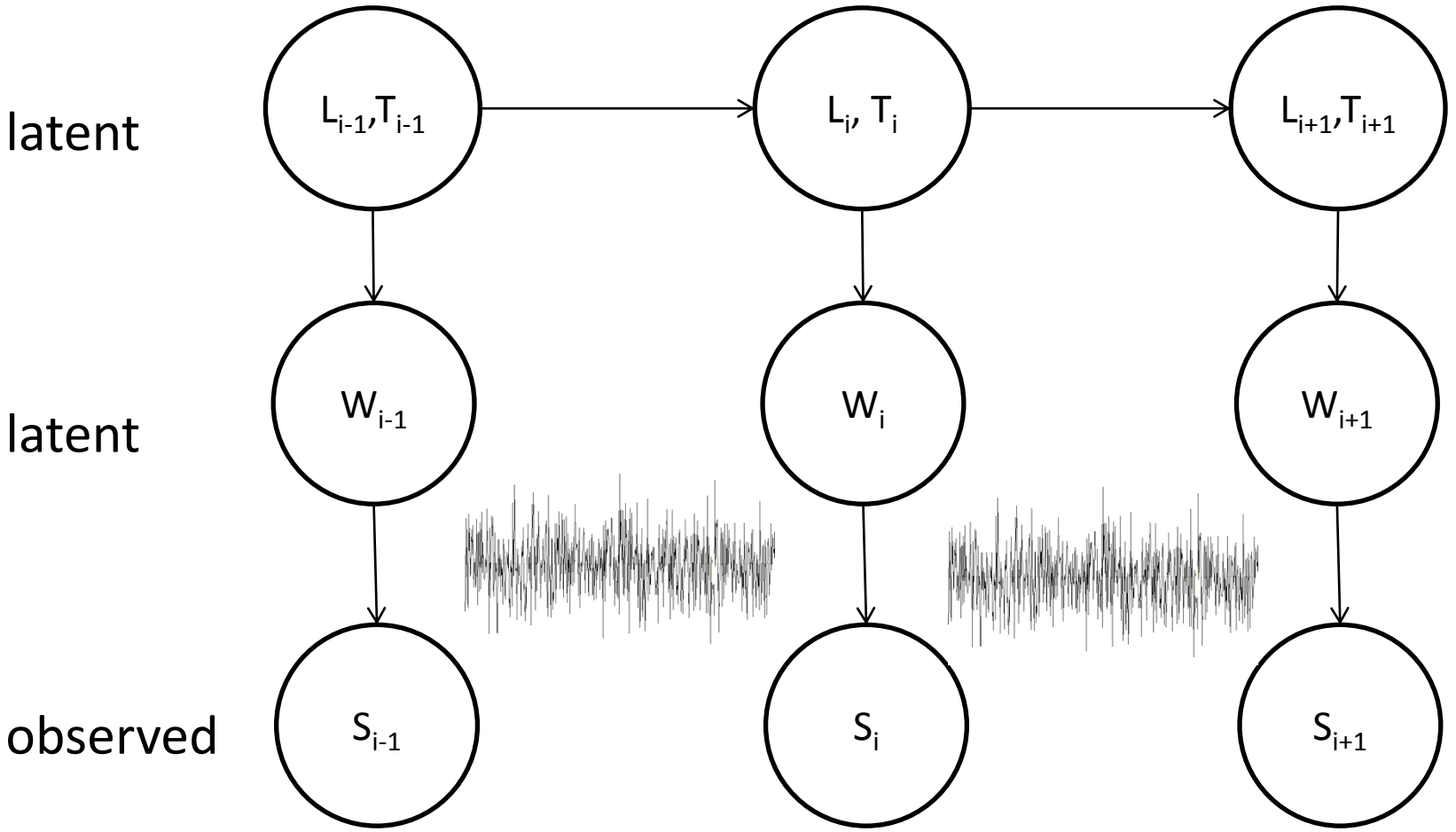
Marginalized Model

$$\arg\max_{L,T} P(L,T \mid S) =$$

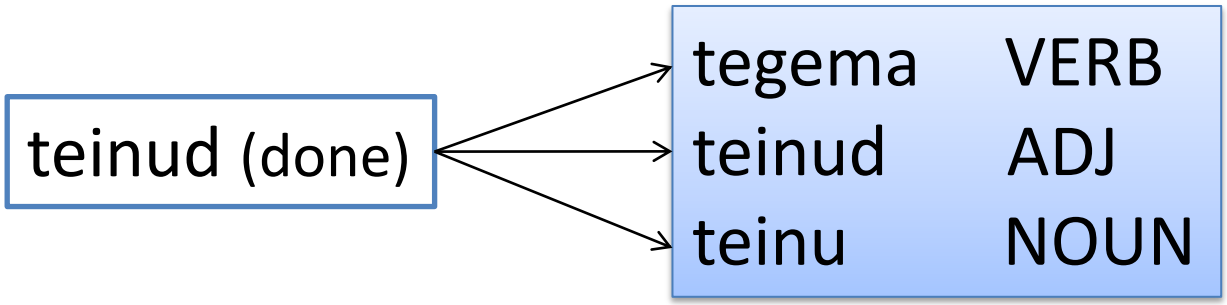
$$\arg\max_{L,T} \sum_{W} P(L,T \mid W) P(W \mid S)$$

W – word
S – spelling
L – lemma
T – POS tag

HMM Marginalized Model



Additional Disambiguation



$$\arg\max_{L,T} P(L,T \mid W) = \arg\max_{L,T} P(W \mid L,T)P(L,T)$$

W – word
L – lemma
T – POS tag

Training

Language models:

- newspaper corpus
- 87,9 mio tokens
- 100000 types

Error model:

- weighted edit distance
- 2780 error-correction pairs
- collected from newspaper texts

Labelled data:

- ca 6500 running words
- 3250 for tuning, 3250 for testing

Results

Model	Accuracy (L+T)
Baseline	91.10
Non-word Pipeline	**92.15**
Non-word Marginalized	90.95
Non-word HMM	90.95
Real-word Pipeline	**92.00**
Real-word Marginalized	90.80
Real-word HMM	89.83

Issues

- Wide variety of errors
- Unusual word order
- Incorrect language usage semanticaly, syntactically

  mina raamastusin, kui ei saanud sõita Veenemaale

  mul oli väga solvavalt, sest mull oli hea pall

  mullu mulle vedus

  meie elu on triibuline

Conclusions

- Experimented with different correction-lemmatisation strategies
- Simplest model works the best
- Spelling correction alone is not enough

Thank you!

Questions?