Bayesian Language Modelling of German Compounds

Jan Botha[*] Chris Dyer[†] Phil Blunsom[*]

[*] Department of Computer Science
University of Oxford

[†] School of Computer Science
Carnegie Mellon University

COLING 2012

Why Statistical Language Modelling?

- Central to tasks where sentences are hypothesised
  - Machine Translation
  - Speech Recognition
  - Optical Character Recognition

Token-based n-gram models

Task: $P\,(\text{Wir fahren mit der Bahn}) = ?$ *We're going by train*

Token-based n-gram models

Task: $P\,(\text{Wir fahren mit der Bahn}) = ?$ *We're going by train*

Markov assumption

Makes parameter estimation feasible

e.g. $P(\text{Bahn} \mid \text{mit der}) = 0.3$

Token-based n-gram models

Task: $P\,(\text{Wir fahren mit der Bahn}) = ?$ *We're going by train*

Markov assumption

Makes parameter estimation feasible

$$\text{e.g. } P(\text{Bahn} \mid \text{mit der}) = 0.3$$

Vocabulary Assumptions

Closed vocabulary, independent words

$$P\left(\begin{array}{c}\text{Bahn}\\ \text{Familie}\end{array} \mid \text{mit der}\right) = ?$$

Token-based n-gram models

Task: $P\,(\text{Wir fahren mit der Bahn}) = ?$ *We're going by train*

Markov assumption

Makes parameter estimation feasible

$$\text{e.g. } P(\text{Bahn} \mid \text{mit der}) = 0.3$$

Vocabulary Assumptions

Closed vocabulary, independent words

$$P\left(\begin{array}{c}\text{Bahn}\\ \text{Familie}\\ \text{Mauer}\end{array} \mid \text{mit der}\right) = ?$$

Token-based n-gram models

Task: $P\,(\text{Wir fahren mit der Bahn}) = ?$ *We're going by train*

Markov assumption

Makes parameter estimation feasible

$$\text{e.g. } P(\text{Bahn} \mid \text{mit der}) = 0.3$$

Vocabulary Assumptions

Closed vocabulary, independent words

$$P\left(\begin{array}{c}\text{Bahn}\\\text{Familie}\\\text{Mauer}\\\text{Straßenbahn}\\\text{U-Bahn}\\\text{Bobbahn}\end{array} \mid \text{mit der}\right) = ?$$

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Highly Productive Process

Regal *shelf*

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Highly Productive Process

Regal	*shelf*
Buchregal	*bookshelf*

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Highly Productive Process

Regal	*shelf*
Buchregal	*bookshelf*
Buchregalhersteller	*bookshelf manufacturer*

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Highly Productive Process

Regal	*shelf*
Buchregal	*bookshelf*
Buchregalhersteller	*bookshelf manufacturer*
Buchregalherstellername	*bookshelf manufacturer's name*
...	

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Regularity

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Regularity

- compound = modifiers & head

Compounding

Example

Wir fahren mit der	Bahn		*train*
	Straßenbahn		*tram*
	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Regularity

- compound = modifiers & head
- head determines syntactic properties

Compounding

Example

	Bahn		*train*
	Straßenbahn		*tram*
Wir fahren mit der	U-bahn	.	*metro*
	Bobbahn		*bobsled*
	Achterbahn		*rollercoaster*

Regularity

- compound = modifiers & head
- head determines syntactic properties
- generalising over modifiers would reduce sparsity

Related Work: Factored Language Model

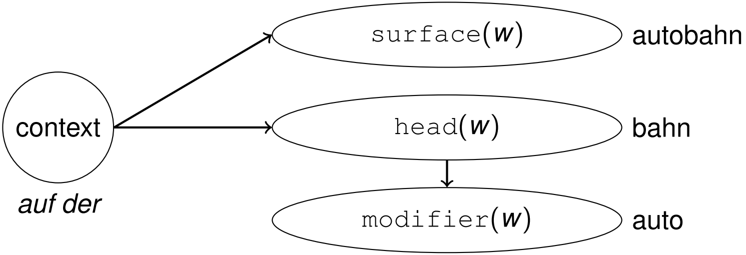
Each token w consists of k features, fixed dependencies

(Bilmes & Kirchhoff, 2003)

Related Work: Factored Language Model

Each token *w* consists of *k* features, fixed dependencies

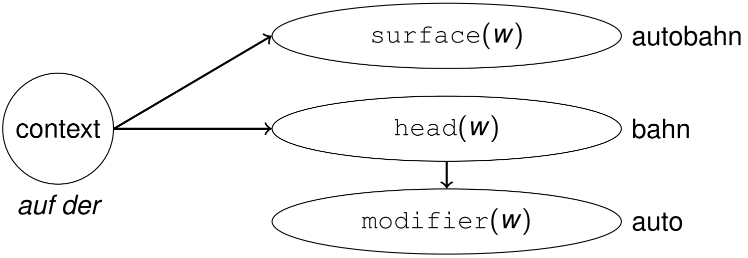
(Bilmes & Kirchhoff, 2003)



Related Work: Factored Language Model

Each token *w* consists of *k* features, fixed dependencies

(Bilmes & Kirchhoff, 2003)



Shortcoming: Compound lengths vary freely

Auto	+	bahn	+	geschwindigkeit	+	s	+	grenze
motor		*way*		*speed*				*limit*

Related Work: Splitting and Merging

Preprocess data: split compounds

(e.g. Koehn & Knight, 2003; Stymne 2008; Macherey et al. 2011)

Related Work: Splitting and Merging

Preprocess data: split compounds

(e.g. Koehn & Knight, 2003; Stymne 2008; Macherey et al. 2011)

Example

		Context	Predicted Token
Standard:	Wir fahren	auf der	Autobahn
Split:	Wir fahren auf	der Auto	bahn

Related Work: Splitting and Merging

Preprocess data: split compounds

(e.g. Koehn & Knight, 2003; Stymne 2008; Macherey et al. 2011)

Example

		Context	Predicted Token
Standard:	Wir fahren	auf der	Autobahn
Split:	Wir fahren auf	der Auto	bahn

Shortcoming: Important sentential context lost

Model Aims

- Condition head on previous words
- Condition compound-internal structure on head

Model Aims

- Condition head on previous words
- Condition compound-internal structure on head

über die grenze

Model Aims

- Condition head on previous words
- Condition compound-internal structure on head

über die geschwindigkeits grenze

Model Aims

- Condition head on previous words
- Condition compound-internal structure on head

über die bahn geschwindigkeits grenze

Model Aims

- Condition head on previous words
- Condition compound-internal structure on head

über die Auto bahn geschwindigkeits grenze

Model Aims

- Condition head on previous words
- Condition compound-internal structure on head

über die Auto bahn geschwindigkeits grenze

- *but* keep full surface form in model

Hierarchical Pitman-Yor Language Model (HPYLM)

Standard back-off smoothing

Hierarchical Pitman-Yor Language Model (HPYLM)

Standard back-off smoothing

mit der Autobahn

$\uparrow$

$P(\text{Autobahn} \mid \text{mit der})$

$P(\text{Autobahn} \mid \ldots\ \text{der})$

$P(\text{Autobahn} \mid \ldots\ \ldots)$

Hierarchical Pitman-Yor Language Model (HPYLM)

Standard back-off smoothing

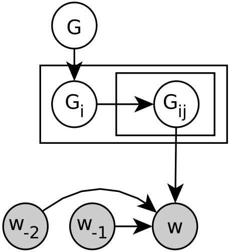
mit der Autobahn

P(Autobahn | mit der)

P(Autobahn | … der)

P(Autobahn | … …)

Hierarchical Prior



Hierarchical Pitman-Yor Language Model (HPYLM)

Standard back-off smoothing

mit der Autobahn

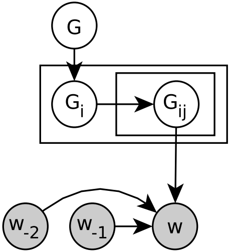
$\uparrow$

$P(\text{Autobahn} \mid \text{mit der})$

$P(\text{Autobahn} \mid \ldots \ \text{der})$

$P(\text{Autobahn} \mid \ldots \ \ldots)$

Hierarchical Prior



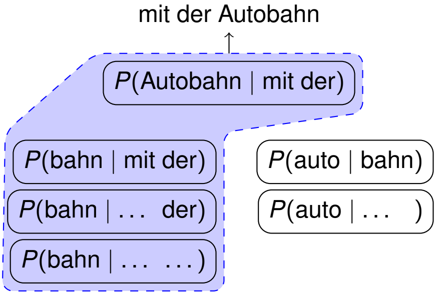
Conditional distributions drawn from Pitman-Yor process:

$$G \sim PY\left(d, \theta, G_0\right)$$

discount strength base distribution

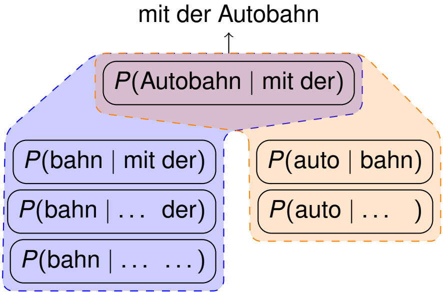
Our Model: HPYLM with Compounds

Proposed back-off



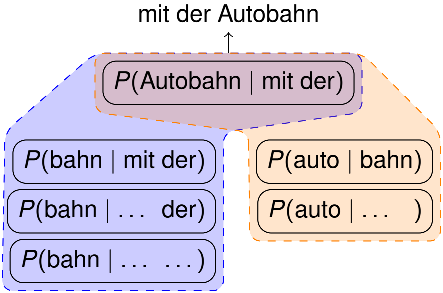
Our Model: HPYLM with Compounds

Proposed back-off

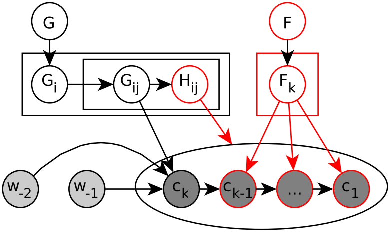


Our Model: HPYLM with Compounds

Proposed back-off



Extended Hierarchical Prior



Overview of Setup

- Data from WMT '11 shared task
- Language Models
  4-gram, unless otherwise stated
  trained on 59m tokens (Europarl, news & commentary)
- English→German Translation System
  trained on 1.7m parallel sentences (Europarl)

Overview of Setup

- Data from WMT '11 shared task
- Language Models
  4-gram, unless otherwise stated
  trained on 59m tokens (Europarl, news & commentary)
- English→German Translation System
  trained on 1.7m parallel sentences (Europarl)
- Vocabulary sparsity – ratio of English/German:
  baseline: 3.13
  decompounded: 1.36

Comparison of Methods for Predicting Unseen Text

	Perplexity
Modified Kneser-Ney	299.9
HPYLM	294.1

Lower is better.

Comparison of Methods for Predicting Unseen Text

	Perplexity
Modified Kneser-Ney	299.9
HPYLM	294.1
Our model	**293.6**

Lower is better.

proposed: über die Auto bahn geschwindigkeits grenze

Comparison of Methods for Predicting Unseen Text

	Perplexity
Modified Kneser-Ney	299.9
HPYLM	294.1
Our model	**293.6**
Our model (inverted)	305.5

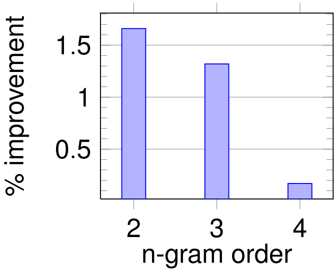
Lower is better.

proposed: über die Auto bahn geschwindigkeits grenze

inverted: über die Auto bahn geschwindigkeits grenze

Effect of Scaling Context Size

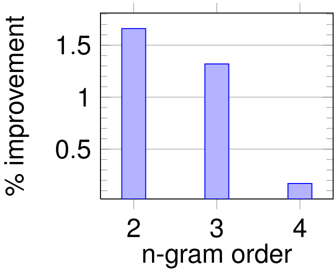
% Relative improvement over HPYLM baseline

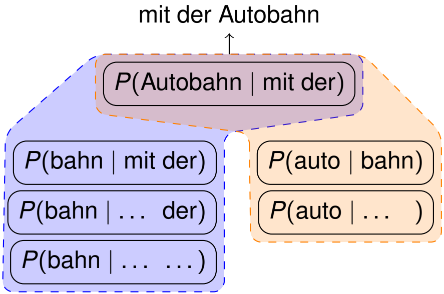


Higher is better.

Effect of Scaling Context Size

% Relative improvement over HPYLM baseline





Higher is better.

Missing: $P(\text{Autobahn} \mid \ldots \text{der})$

Machine Translation Results

English→German

	BLEU
Modified Kneser-Ney	13.9
HPYLM	13.9
Our model	13.9
Our model (inverted)	13.7

Higher BLEU is better.

Machine Translation Results

English→German

only 3.7% of reference tokens are compounds

	BLEU
Modified Kneser-Ney	13.9
HPYLM	13.9
Our model	13.9
Our model (inverted)	13.7

Higher BLEU is better.

Compound Translation Accuracy

Compare compounds in output against reference sentences

	Precision	**Recall**	**F-score**
Modified Kneser-Ney	25.4	17.1	20.5
HPYLM	24.3	**17.5**	20.4
Our model	**27.5**	17.3	**21.3**
Our model (inverted)	23.7	17.2	19.9

Higher is better.

Summary

- Productive compounding is an important source of sparsity.
- Proposed n-gram language model that accounts for productive compounding.
- Improved accuracy of translated compounds, while matching baseline BLEU score

Summary

- Productive compounding is an important source of sparsity.
- Proposed n-gram language model that accounts for productive compounding.
- Improved accuracy of translated compounds, while matching baseline BLEU score

- Future Work
  - Evaluate with an MT system that outputs *novel* compounds
  - Model compounds that occur within context

Thank you.

	BLEU
Modified Kneser-Ney	13.9
HPYLM	13.9
Our model	13.9

	LM-training	BLEU
Modified Kneser-Ney	< 1 hour	13.9
HPYLM	3 days	13.9
Our model	6 days	13.9

	LM-training	BLEU
Modified Kneser-Ney	$<$ 1 hour	13.9
HPYLM	3 days	13.9
Our model	6 days	13.9
Our model (fastapprox)	4 hours	13.6