Translation to morphologically rich languages

Alexander Fraser

LMU Munich

January 17th 2019

Introduction

- Translation to morphologically rich languages is difficult
- Hypothesis: if we are willing to do language-specific engineering, we can do better
- Before: DFG (German National Science Foundation), FP7 "TTC" project, H2020 Health in my Language "HimL"
- Now: ERC Starting Grant
- **Basic research question**: can we integrate linguistic resources for morphology and syntax into (large scale) statistical machine translation and achieve better generalization?
- Will talk about translation from German to English briefly
- Primary focus: translation to German (and more recently, Czech)
  - Critical issue: how to generate morphology (for German or Czech) which is more specified than in the source language (English)?

Outline

1. Linguistically-rich pipelines for PBSMT
2. Integrating morphological generation into PBSMT
3. Target-side segmentation for neural machine translation
4. Generalizing over stems in neural machine translation

Translating from English to German

Linguistically rich pipeline (example from Cap on next slide, credit also to Ramm and Weller Di Marco):

- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, **tense**, aspect
- Use phrase-based model to translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features (e.g., **case**, number, gender)
  - Case is particularly difficult (see, e.g., Weller ACL 2013 for more examples)

Translating from English to German

Linguistically rich pipeline (example from Cap on next slide, credit also to Ramm and Weller Di Marco):

- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, **tense**, aspect
- Use phrase-based model to translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features (e.g., **case**, number, gender)
  - Case is particularly difficult (see, e.g., Weller ACL 2013 for more examples)

Translating from English to German

Linguistically rich pipeline (example from Cap on next slide, credit also to Ramm and Weller Di Marco):

- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, **tense**, aspect
- Use phrase-based model to translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features (e.g., **case**, number, gender)
  - Case is particularly difficult (see, e.g., Weller ACL 2013 for more examples)

Translating from English to German

Linguistically rich pipeline (example from Cap on next slide, credit also to Ramm and Weller Di Marco):

- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, **tense**, aspect
- Use phrase-based model to translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features (e.g., **case**, number, gender)
  - Case is particularly difficult (see, e.g., Weller ACL 2013 for more examples)

Translating from English to German

Linguistically rich pipeline (example from Cap on next slide, credit also to Ramm and Weller Di Marco):

- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, **tense**, aspect
- Use phrase-based model to translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features (e.g., **case**, number, gender)
  - Case is particularly difficult (see, e.g., Weller ACL 2013 for more examples)

Translating from English to German

Linguistically rich pipeline (example from Cap on next slide, credit also to Ramm and Weller Di Marco):

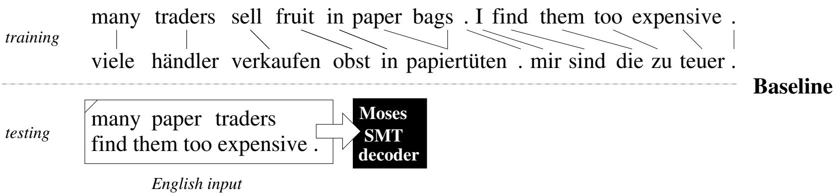
- Use classifiers to classify English clauses with their German word order
- Predict German verbal features like person, number, **tense**, aspect
- Use phrase-based model to translate English words to German lemmas (with split compounds)
- Create compounds by merging adjacent lemmas
  - Use a sequence classifier to decide where and how to merge lemmas to create compounds
- Determine how to inflect German noun phrases (and prepositional phrases)
  - Use a sequence classifier to predict nominal features (e.g., **case**, number, gender)
  - Case is particularly difficult (see, e.g., Weller ACL 2013 for more examples)

Compound Merging and Inflection Prediction

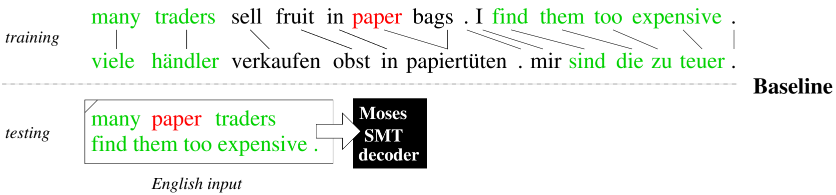
Compound Merging and Inflection Prediction

	many traders sell fruit in paper bags . I find them too expensive .
training	
	viele händler verkaufen obst in papiertüten . mir sind die zu teuer .

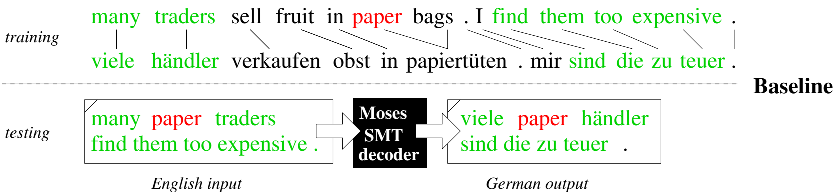
Compound Merging and Inflection Prediction



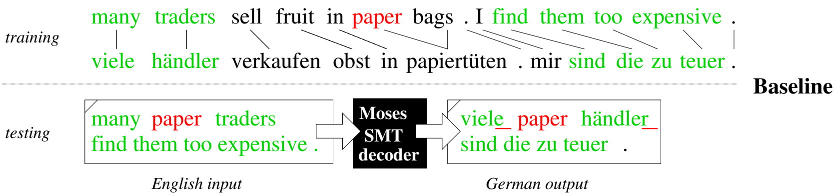
Compound Merging and Inflection Prediction



Compound Merging and Inflection Prediction



Compound Merging and Inflection Prediction



Compound Merging and Inflection Prediction

training

many traders sell fruit in paper bags . I find them too expensive .

viele händler verkaufen obst in papiertüten . mir sind die zu teuer .

Baseline

testing

many paper traders find them too expensive .

Moses SMT decoder

viele_ paper händler_ sind die zu teuer .

English input *German output*

---

training

many traders sell fruit in paper bags . I find them too expensive .

viele händler verkaufen obst in papiertüten . mir sind die zu teuer .

Our system

Compound Merging and Inflection Prediction

training

many traders sell fruit in paper bags . I find them too expensive .

viele händler verkaufen obst in papiertüten . mir sind die zu teuer .

Baseline

testing

many paper traders
find them too expensive .

English input

Moses SMT decoder

viele_ paper händler_
sind die zu teuer .

German output

---

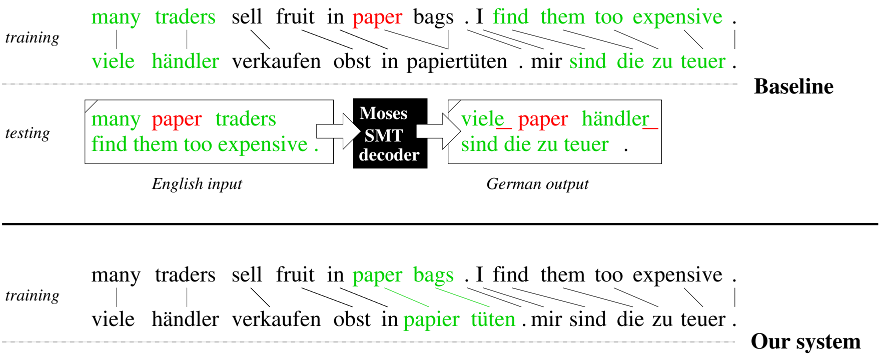
training

many traders sell fruit in paper bags . I find them too expensive .

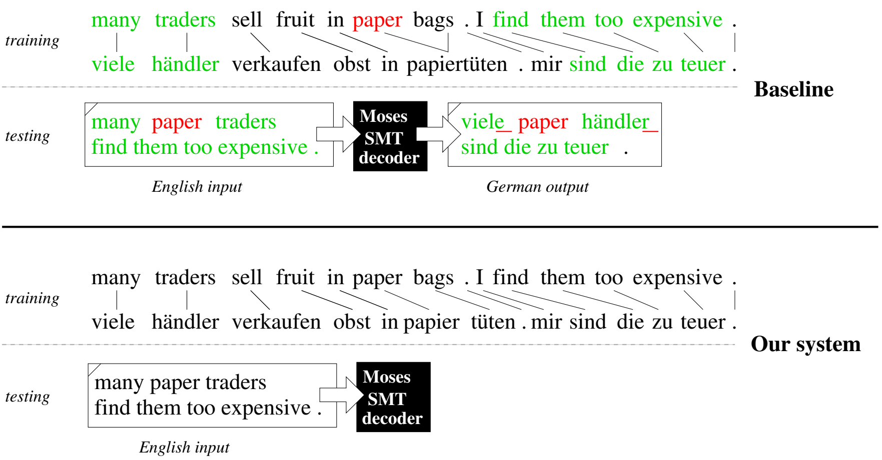
viele händler verkaufen obst in papiertüten . mir sind die zu teuer .

Our system

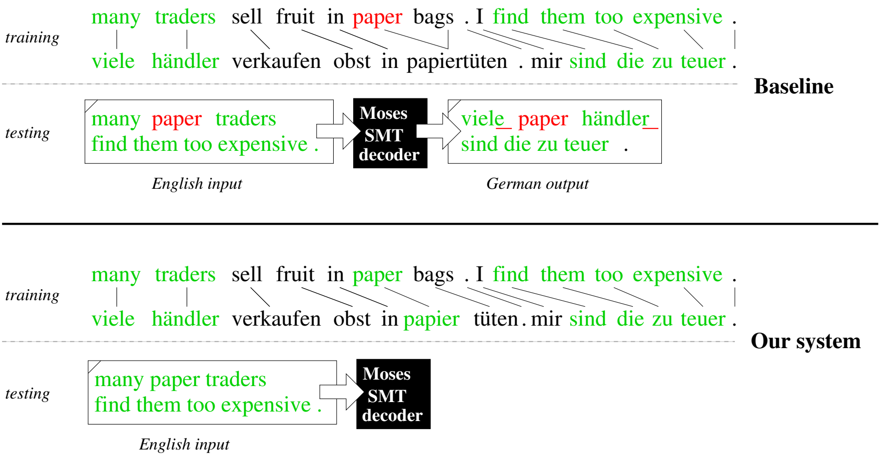
Compound Merging and Inflection Prediction



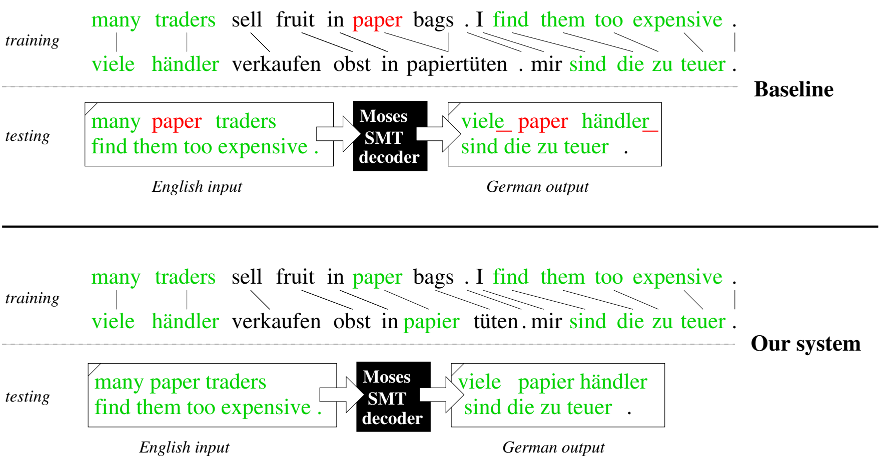
Compound Merging and Inflection Prediction



Compound Merging and Inflection Prediction

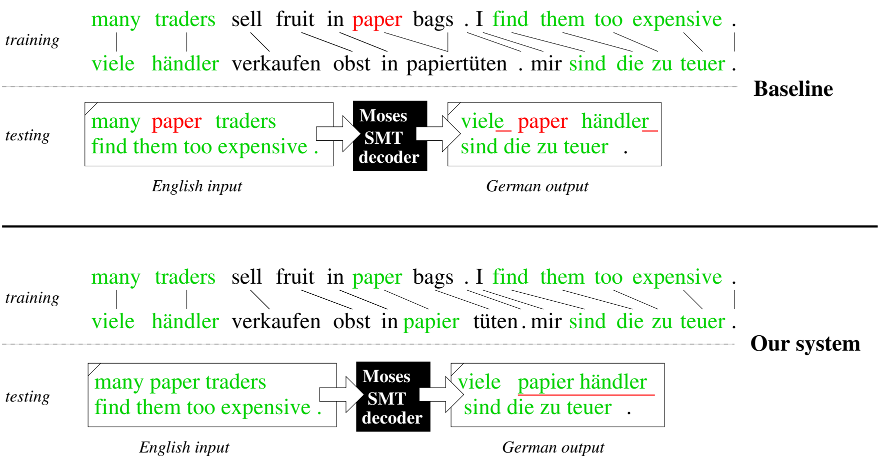


Compound Merging and Inflection Prediction



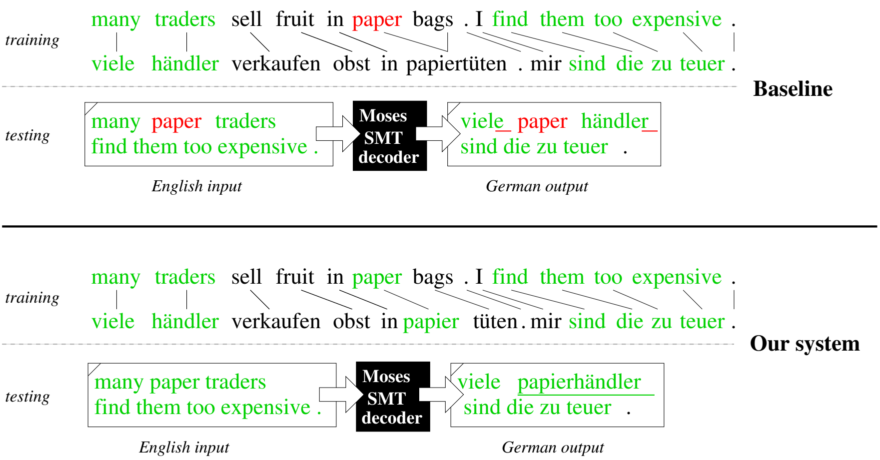
Translate into split and lemmatised German

Compound Merging and Inflection Prediction



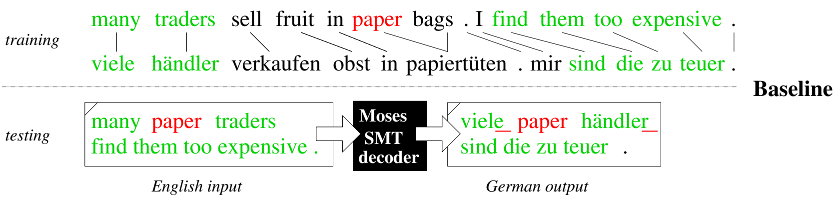
Step 1: Compound Merging

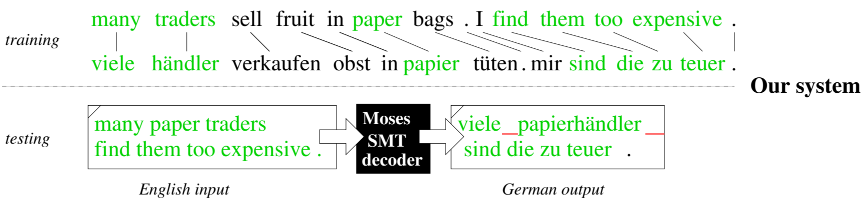
Compound Merging and Inflection Prediction



Step 1: Compound Merging

Compound Merging and Inflection Prediction





Step 1: Compound Merging   Step 2: Re-Inflection

Compound Merging and Inflection Prediction

Step 1: Compound Merging   Step 2: Re-Inflection

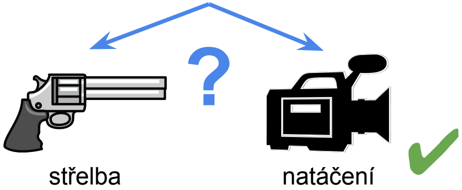
Outline

Integrating Morphological Generation into Moses

- We integrated discriminative models for morphological richness into Moses
- Based on “VW”, Vowpal Wabbit (Langford and others)
- Similar to previous work on integrated discriminative lexicon models, but different:
  - Modeling target-side in online decoding
  - Synthesizing missing surface forms for known lemmas
- Papers: Tamchyna et al. Target-side Context for Discriminative Models in Statistical Machine Translation. ACL 2016.
- Huck et al. Producing Unseen Morphological Variants in Statistical Machine Translation. EACL 2017.

Why Context Matters in MT: Source

shooting of the expensive film



Wider **source** context required for disambiguation of word **sense.**

Previous work has looked at using source context in MT.

Why Context Matters in MT: Target

the man | saw | a cat | .

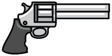
si všiml
uviděl

kočka	*nominative*
kočky	*genitive*
kočce	*dative*
kočku	*accusative*
kočko	*vocative*
kočce	*locative*
kočkou	*instrumental*

Correct case depends on how we translate the previous words.

Wider **target** context required for disambiguation of word **inflection**.

How Does PBMT Fare?

shooting of the **film** .	**natáčení** filmu .	✔	
shooting of the expensive **film** .	**střelby** na drahý film .	✘	

the man saw **a cat** .	muž uviděl **kočku**$_{acc}$.	✔
the man saw **a black cat** .	muž spatřil **černou**$_{acc}$ **kočku**$_{acc}$.	✔
the man saw **a yellowish cat** .	muž spatřil **nažloutlá**$_{nom}$ **kočka**$_{nom}$.	✘

Morphological Variants

Translating into morphologically rich languages is difficult.

- Permissible morphological variants remain unseen in training
- SMT systems fail at producing them

Our approach: Providing full coverage of morphological variants.

- Missing variants are synthesized
- The decoder can choose freely amongst all inflected forms

Challenge: How to score unseen morphological variants?

- Additional features in a phrase-based model
- A discriminative classifier that is designed to generalize to unseen morphological variants

Motivation: English→Czech

case	surface form	50K	500K	5M	50M
1	čéšky	•	•	•	•
2	čéšek	–	•	•	•
3	čéškám	–	–	•	•
4	čéšky	○	○	•	•
5	čéšky	○	○	○	○
6	čéškách	–	•	•	•
7	čéškami	–	–	–	•

Morphological variants of the Czech lemma "čéška". For differently sized corpora (50K/500K/5M/50M), "•" indicates that the variant is present, and "○" that the same surface form realization occurs, but in a different syntactic case.

Generating Unseen Variants

baseline phrase		**lemma**		**morphological generator**
kneecaps ||| čéšky	$\rightarrow$	čéška	$\rightarrow$	čéšky čéšek čéškám čéšky čéšky čéškách čéškami

Missing variants are added as new translation options.

Settings:

- **word** Phrases of length 1 on both source and target side
- **mtu** Arbitrary length of the phrase source side
- $\star$ Force some attributes to match the original word form ("tag template", e.g. for number, tense, . . .)

Czech morphological generator: **MorphoDiTa** (Straková et al., 2014)

Discriminative Classifier

morph-vw (*Vowpal Wabbit for Morphology*):

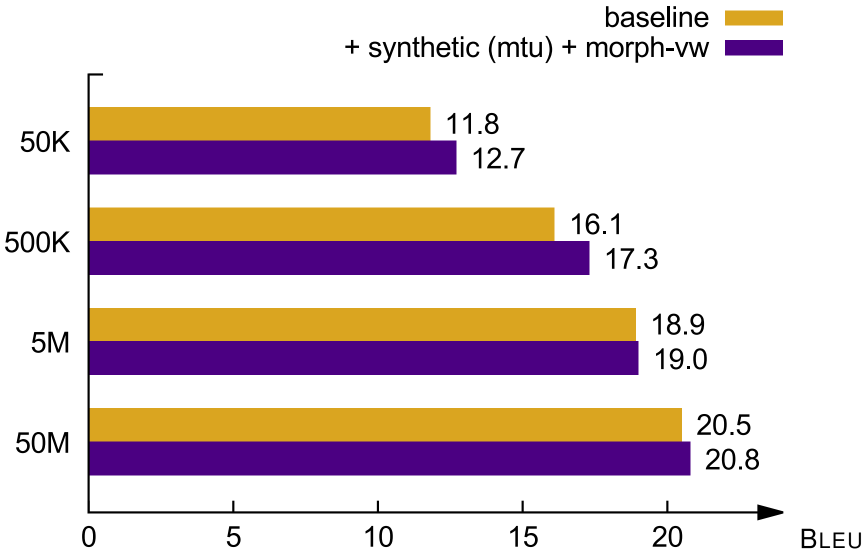
- A decoder-integrated classifier that generalizes to unseen morphological variants.

Feature templates:

feature type	configurations
source indicator	l, t
source internal	l, l+r, l+p, t, r+p
source context	l (-3,3), t (-5,5)
target indicator	l, t
target internal	l, t
target context	l (-2), t (-2)

l (lemma), **t** (morphosyntactic tag), **r** (syntactic role), **p** (lemma of dependency parent). Numbers in parentheses indicate context size.

Experimental Results: Scaling Up/Down



Translation Example (500K)

input:

now , six in 10 Republicans have a favorable view of Donald Trump **.**

baseline:

teď , šest **v** 10 **republikáni** mají příznivý **výhled Donald Trump .**

*now, six **in**$_{location}$ 10 **Republicans**$_{nom}$ have a_favorable **outlook Donald**$_{nom}$ **Trump**$_{nom}$.*

+ synthetic (mtu) + morph-vw:

teď , šest **do** deseti republikánů má příznivý názor na Donalda Trumpa **.**

now, six **into** *ten$_{gen}$ Republicans$_{gen}$ have a_favorable opinion of Donald$_{acc}$ Trump$_{acc}$.*

Summary

Full coverage of all valid inflected target word forms

- for each known lemma
- with direct integration into phrase-based search

Effective scoring of unseen morphological variants

- utilizing a tailored discriminative classifier
- taking both source and target context into account

Substantial BLEU score improvements,
particularly on small to medium resource translation tasks

Treating the same phonemena in neural machine translation (NMT)

Quick summary:

- We've done work showing that word ordering is not an issue for NMT (e.g., Ramm EAMT 2016).
  - For instance, Martin mentioned German particle verbs, a problem in phrase-based SMT. In NMT they seem to be fine.
- For morphological generalization, working at the character-level (character NMT) is promising, a few papers, but this isn't working yet
- It has been claimed that simple frequency-based approaches to segmentation generalize well over phenomena such German inflection and compounds
- I think using linguistic knowledge we should be able to beat these approaches, I will present two pieces of work

Outline

Motivation for target-side segmentation

Limited-size set of target symbols in NMT

- Large vocabularies not tractable in practice
- BPE word segmentation now in common use

Can NMT learn linguistic processes of word formation?

- Morphological variation
- Compounding
- BPE not tailored for this

Use linguistically informed target word segmentation!

- For better vocabulary reduction
- For reduction of data sparsity
- For better open vocabulary translation
- Here: English$\rightarrow$German

Word Segmentation Strategies

1. **Suffix splitting**
   - Snowball stemmer adapted for word segmentation
2. **Prefix splitting**
   - A simple in-house script that covers common German prefixes
3. **Compound splitting**
   - Frequency-based splitter as implemented in Moses
4. **Byte pair encoding (BPE)**
   - With 50K merge operations
5. **Cascading of segmenters**
   - Pipelined application of the different techniques

Suffixes

Conjugation	singular		plural	
person	**English**	**German**	**English**	**German**
1st	I learn	ich lern**e**	we learn	wir lern**en**

Declension	singular		plural	
case	**English**	**German**	**English**	**German**
nominative	the dog	der Hund	the dogs	die Hund**e**
accusative	the dog	den Hund	the dogs	die Hund**e**

Nominalization (-heit, -keit, -nis, -ung, . . .)
Adjective derivation (-ig, -isch, -sam, . . .)
Present participle (-end)

Prefixes

Formation of new words, often involving semantic change

German	English
bringen	to bring
mitbringen	to bring along
wegbringen	to bring away
anbringen	to apply; to install
umbringen	to kill
ver-/zubringen	to spend (time)
elegant	elegant
hochelegant	highly elegant
überelegant	overly elegant
unelegant	inelegant

German Affixes in Word Segmentation

Suffixes
-e, *-em*, *-en*, *-end*, *-enheit*, *-enlich*, *-er*, *-erheit*, *-erlich*, *-ern*, *-es*, *-est*, *-heit*, *-ig*, *-igend*, *-igkeit*, *-igung*, *-ik*, *-isch*, *-keit*, *-lich*, *-lichkeit*, *-s*, *-se*, *-sen*, *-ses*, *-st*, *-ung*
Prefixes (excerpt)
ab-, *an-*, *anti-*, *auf-*, *aus-*, *auseinander-*, *außer-*, *be-*, *bei-*, *binnen-*, *bitter-*, *blut-*, *brand-*, *dar-*, *des-*, *dis-*, *durch-*, *ein-*, *empor-*, *endo-*, *ent-*, *entgegen-*, . . . (+ 112 more)

BPE, Cascading, Reversibility

BPE

- Reduces the number of distinct symbols to approximately a configurable amount.

Cascading of Segmenters

- Combine suffix splitting, prefix splitting, compound splitting, and BPE by applying them after another.

Reversibility

- Special markers for desegmentation in postprocessing.

Word Segmentations: Example

English	*they all mail deliberately deceptive documents to small businesses across Europe .*
BPE	*sie alle versch ## icken vorsätzlich irreführende Dokumente an Kleinunternehmen in ganz Europa .*
compound + BPE	*sie alle verschicken vorsätzlich #L irre @@ führende Dokumente an #U klein @@ unter @@ nehmen in ganz Europa .*
suffix + BPE	*sie all $$e verschick $$en vorsätz $$lich irreführ $$end $$e Dokument $$e an Kleinunternehm $$en in ganz Europa .*
suffix + compound + BPE	*sie all $$e verschick $$en vorsätz $$lich #L Irre @@ führ $$end $$e Dokument $$e an #U klein @@ Unternehm $$en in ganz Europa .*
suffix + prefix + compound + BPE	*sie all $$e ver§§ schick $$en vor§§ sätz $$lich #L Irre @@ führ $$end $$e Dokument $$e an #U klein @@ Unternehm $$en in ganz Europa .*

Europarl: English→German NMT

System	test2007		test2008	
	BLEU	TER	BLEU	TER
BPE	25.8	60.7	25.6	60.9
compound + BPE	25.9	60.3	25.5	60.6
suffix + BPE	**26.3**	60.0	**26.0**	**60.1**
suffix + compound + BPE	26.2	**59.8**	25.8	60.2
suffix + prefix + compound + BPE	26.1	**59.8**	25.9	60.6

Pros & Cons

Pros:

- Simple and (with basic linguistic knowledge) easy to implement for many languages.
- No full-fledged morphological analyzer or generator required.
- Good enough to improve translation quality.

Cons:

- Plain segmentation is less powerful than proper morphological analysis. (Consider irregular verbs, for instance.)
- Lack of interplay between source and target segmentation. (E.g.: kind**ness** – Freundlich**keit**; rest**less** – rast**los**; bookshelf – Bücher|regal vs. reading corner – Bücher|ecke)

Summary

- Linguistically motivated target-side word segmentation improves NMT into an inflected and compounding language (up to $+0.5$ BLEU and -0.9 TER).
- Linguistic word formation processes can be learned from the segmented data.
- Top system in WMT 2017 en-de news according to human evaluation (but not according to BLEU!)

Outline

1. Linguistically-rich pipelines for PBSMT
2. Integrating morphological generation into PBSMT
3. Target-side segmentation for neural machine translation
4. Generalizing over stems in neural machine translation

Motivation: Morphology in Current NMT Systems

- subwords (BPE, wordpieces) are typically used to handle large target-side vocabularies
- no explicit connection between different surface forms of a single lemma
    - BPE does not reliably split at morpheme boundaries
    - BPE can only handle purely concatenative word formation, but:
        - Umlaut in German: "***Haus***", "***Häus***-*er*"
        - root-internal changes in Czech: *"**dům**, **dom**-y", "**rak**-a, **rac**-i"*
- no information about morphological features of target-side words
- no systematic way of generating unseen forms of known lemmas

Two-Step Translation

1. train an NMT model to translate input surface forms into a sequence of morphological **tags** and **lemmas**:

   input: there are a million different kinds of pizza .

   target: VB-P---3P existovat NNIP1---- milión NNIP2---- druh NNFS2---- pizza Z:------- .

2. use a deterministic morphological generator to produce the final output:

   final: existují miliony druhů pizzy .

Two-Step Translation

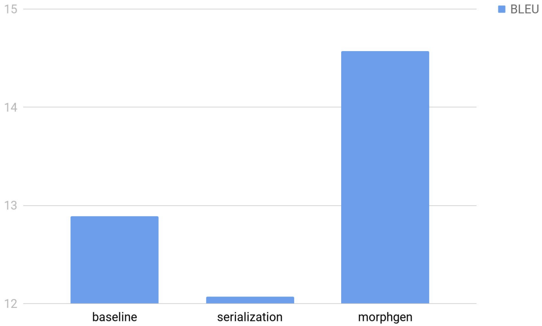
- unlike BPE, provides a linguistically justified generalization
  - using lemmas leads to a reduction in vocabulary size
  - more robust statistical estimates for lexical mapping
- morphological information is explicit
  - allows the network to learn and use a “morphological LM” as a by-product
- allows to generate novel word forms (of known lemmas)
  - simply produce a new pair (tag, lemma)
- inspired by the *serialization* approach in Nadejde et al. 2017: Syntax-aware Neural Machine Translation Using CCG
  - morphological tags instead of CCG supertags
  - lemmas instead of surface forms

Experiments: English-Czech

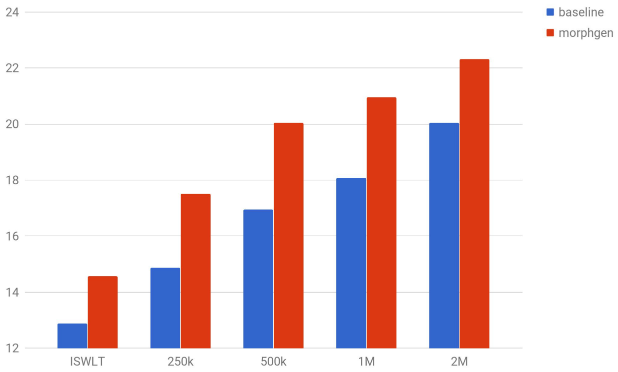
- MorphoDiTa for tagging, lemmatization and generation
- data:
  - dev/test: IWSLT test sets 2012, 2013
  - training: WIT3 corpus (114k sentence pairs)
  - larger experiments: mix in parts of CzEng 1.0 up to 2 million sentence pairs in total
- all systems built using Nematus
  - 49500 BPE splits
  - embedding size 500, hidden layer size 1024, no drop-out, Adam, max. length 50 (100)
- system variants:
  - **baseline:** ... kinds of pizza . → ... druhů pizzy .
  - **serialization:** ... kinds of pizza . → ... NNIP2---- druhů NNFS2---- pizzy Z:------- . *(forms)*
  - **morphgen:** ... kinds of pizza . → ... NNIP2---- druh NNFS2---- pizza Z:------- . *(lemmas)*

Results: English-Czech

- IWSLT data (114k sents), best system on dev from 3 training runs



English-Czech: Scaling to Larger Data



English-German Translation

- morphological analysis and generation: SMOR
- parsing: BitPar
- example of lemmatized representation:

  trifft→ treffen <+V><3><Sg><Pres><Ind>

- different morphological categories depending on PoS
- in addition to inflection, handle compounding (based on SMOR analysis):

  Meeresboden → Meer $$<NN>$$ Boden <+NN><Masc><Dat><Sg><St>

Experiments: English-German

- data:
  - dev/test: IWSLT test sets 2012, 2013, 2014, training: WIT3 corpus (184k sentence pairs)
  - larger experiments: select parts of WMT17 training corpora up to 500k sentence pairs; dev/test: newstest 2015, 2016
- system parameters
  - 29500 BPE splits
  - embedding size 500, hidden layer size 1024, drop-out enabled, Adam, max. length 50 (100)
- system variants:
  - **baseline:**

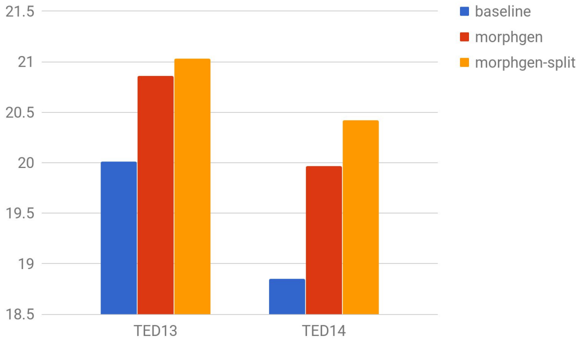
    ... sea floor ... → ... Meeresboden ...
  - **morphgen:**

    ... sea NN floor NN ... → ... Meer<NN>Boden <+NN><Masc><Dat><Sg><St> ...
  - **morphgen-split:**

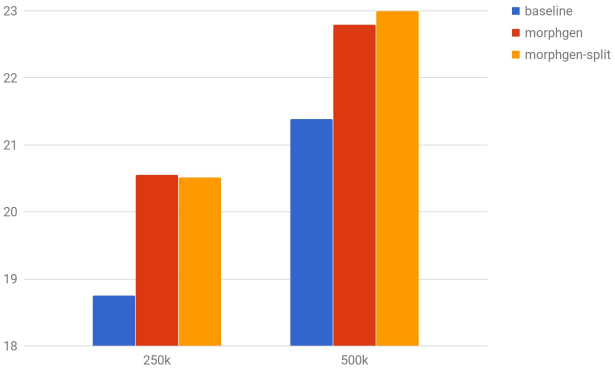
    ... sea NN floor NN ... → ... Meer $$<NN>$$ Boden <+NN><Masc><Dat><Sg><St> ...

Results: English-German

- IWSLT data, average of 2 training runs



English-German: Scaling to Larger Data



Analysis & Discussion

- the network learns to produce well-formed outputs perfectly
  - tags and lemmas always interleave
- combinations of lemma and tag *almost* always valid
  - only a handful of errors
  - morphological generator fails → simply output the lemma
- surprisingly few novel word forms
  - 125 for Czech, 261 for German
  - only a minority of generated forms confirmed by the reference (14 and 21)
  - German: mostly compounds
  - optimization towards cross-entropy most likely discourages novel combinations

Further work

- Using this tag-lemma representation allows us to generalize:
- We have very new work on using **bilingual word embeddings** to mine the translations of out-of-vocabulary words (see papers by Braune, Hangya)
- We are trying to generate new surface forms in translations using a mined lemma translation
- Less successfully (so far), we are also trying to generate novel composita
- Finally, we are also working on modeling extra-sentential context (Stojanovski) and hope to use this representation effectively there as well, e.g., for translating pronouns (consider English “it”)

Summary

- The key questions in data-driven machine translation are about learning from data. This involves linguistic generalization.
    - Discussed two approaches to morphological generation in SMT (pipeline and VW)
    - Presented linguistically-aware segmentation for NMT to German
    - Also showed generalization over word stems for NMT to German and Czech
    - We need better linguistic resources and **more people** working on these problems!
- Credits, my group: Fabienne Braune, Fabienne Cap, Nadir Durrani, Viktor Hangya, Matthias Huck, Anita Ramm, Hassan Sajjad, Dario Stojanovski, Ales Tamchyna, Marion Weller-Di Marco
- Credits, collaborators: Helmut Schmid, Sabine Schulte im Walde, Hinrich Schütze, Ondrej Bojar, Aoife Cahill, Hal Daume, Marcin Junczys-Dowmunt

Questions?

Thank you for your attention

Alexander Fraser

fraser@cis.lmu.de

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 644402 (*HimL*). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).