Linguistic Information in Machine Translation

Marion Weller-Di Marco

dimarco@cis.uni-muenchen.de

June 16 2020

Outline

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern
,	sah
the	der
man	Mann
saw	ein
a	blaues
blue	Auto
car	

the	der
man	Mann
bought	kaufte
a	eine
newspaper	Zeitung

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern
,	sah
the	der
man	Mann
saw	ein
a	blaues
blue	Auto
car	

the	der
man	Mann
bought	kaufte
a	eine
newspaper	Zeitung

- Translate an input sentence given the data in the training corpus:

  the man bought a car

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern	the	der
,	sah	man	Mann
the	der	bought	kaufte
man	Mann	a	eine
saw	ein	newspaper	Zeitung
a	blaues		
blue	Auto		
car			

- Translate an input sentence given the data in the training corpus:

  the man bought a car

  der

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern
,	sah
the	der
man	Mann
saw	ein
a	blaues
blue	Auto
car	

the	der
man	Mann
bought	kaufte
a	eine
newspaper	Zeitung

- Translate an input sentence given the data in the training corpus:

  the man bought a car

  der Mann

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern
,	sah
the	der
man	Mann
saw	ein
a	blaues
blue	Auto
car	

the	der
man	Mann
bought	kaufte
a	eine
newspaper	Zeitung

- Translate an input sentence given the data in the training corpus:

  the man bought a car

  der Mann kaufte

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern
,	sah
the	der
man	Mann
saw	ein
a	blaues
blue	Auto
car	

the	der
man	Mann
bought	kaufte
a	eine
newspaper	Zeitung

- Translate an input sentence given the data in the training corpus:

  the man bought a car

  der Mann kaufte ein

Why using linguistic information in MT?

- MT systems are learnt from (word-aligned) parallel corpora

yesterday	gestern
,	sah
the	der
man	Mann
saw	ein
a	blaues
blue	Auto
car	

the	der
man	Mann
bought	kaufte
a	eine
newspaper	Zeitung

- Translate an input sentence given the data in the training corpus:

  the man bought a car

  der Mann kaufte ein Auto

Why using linguistic information?

- Deriving translation systems from parallel corpora:
  how to translate observed words/phrases in observed contexts

Why using linguistic information?

- Deriving translation systems from parallel corpora:
  how to translate observed words/phrases in observed contexts

- Lack of generalization:
  - to be translated, the exact word needs to be observed
  - requires a lot of training data
  - but: we might have observed a related word or context,
    how to exploit this?

Why using linguistic information?

- Deriving translation systems from parallel corpora:
  how to translate observed words/phrases in observed contexts

- Lack of generalization:
  - to be translated, the exact word needs to be observed
  - requires a lot of training data
  - but: we might have observed a related word or context,
    how to exploit this?

- Amount of available training data
  - what about under-resourced languages or domains?
  - how to make better use of limited data?

Why using linguistic information?

- Deriving translation systems from parallel corpora:
  how to translate observed words/phrases in observed contexts

- Lack of generalization:
  - to be translated, the exact word needs to be observed
  - requires a lot of training data
  - but: we might have observed a related word or context,
    how to exploit this?

- Amount of available training data
  - what about under-resourced languages or domains?
  - how to make better use of limited data?

- Linguistic information to help generalize and to introduce knowledge that is not directly accessible

Why using linguistic information?

- Differences between source and target language
  can make it difficult to learn good translation models

Why using linguistic information?

- Differences between source and target language
  can make it difficult to learn good translation models

- Languages use different mechanisms to encode information,
  for example
  - morphology: varying degrees of complexity
  - syntax: free constituent order vs. strictly configurational

    German: subject/object are defined via grammatical case
    English: subject/object are defined via position in the sentence

Why using linguistic information?

- Differences between source and target language
  can make it difficult to learn good translation models
- Languages use different mechanisms to encode information,
  for example
  - morphology: varying degrees of complexity
  - syntax: free constituent order vs. strictly configurational

    German: subject/object are defined via grammatical case
    English: subject/object are defined via position in the sentence
- Morphology: morphological complexity is challenging in NLP
- Syntax: long distance dependencies or attachment ambiguities

Why using linguistic information?

- Differences between source and target language
  can make it difficult to learn good translation models

- Languages use different mechanisms to encode information,
  for example
  - morphology: varying degrees of complexity
  - syntax: free constituent order vs. strictly configurational

    German: subject/object are defined via grammatical case
    English: subject/object are defined via position in the sentence

- Morphology: morphological complexity is challenging in NLP
- Syntax: long distance dependencies or attachment ambiguities

- Linguistic information to model relevant information

Outline

Morphological complexity across languages

Example: comparing nominal inflection features

- English:

number (only expressed in nouns)
the small dog
the small dogs

Example: comparing nominal inflection features

- English:

number (only expressed in nouns)
the small dog *the small dogs*

- German:

number, gender, case, strong/weak inflection (expressed through the entire phrase)
der kleine Hund *ein kleiner Hund* *dem kleinen Hund* *die kleinen Hunde* *den kleinen Hunden* ...

Example: comparing nominal inflection features

- English:

number (only expressed in nouns)
the small dog *the small dogs*

- German:

number, gender, case, strong/weak inflection (expressed through the entire phrase)
der kleine Hund *ein kleiner Hund* *dem kleinen Hund* *die kleinen Hunde* *den kleinen Hunden* ...

⇒ more word forms observed in German corpus

Example: productive word formation

- Compounding (e.g. German)

Abfall	waste
Abfallsortierung	waste sorting
Abfallsortieranlage	waste sorting plant
Abfallsortieranlagenfachmann	waste sorting plant specialist

Example: productive word formation

- Compounding (e.g. German)

Abfall	waste
Abfallsortierung	waste sorting
Abfallsortieranlage	waste sorting plant
Abfallsortieranlagenfachmann	waste sorting plant specialist

- Agglutinative concatenations (e.g. Turkish)

Turkish	English
duy(-mak)	*(to) sense*
duygu	*sensation*
duygusal	*sensitive*
duygusallaş(-mak)	*(to) become sensitive*
duygusallaştırıl(-mak)	*(to) be made sensitive*
duygusallaştırılmış	*the one who has been made sensitive*
duygusallaştırılamamış	*the one who could not have been made sensitive*
duygusallaştırılamamışlardan	*from the ones who could not have been made sensitive*

Example taken from Ataman et al. (2017)

Translating morphologically complex languages

- Morphologically rich languages: large amount of word forms

Translating morphologically complex languages

- Morphologically rich languages: large amount of word forms
- Problematic for machine translation:
  - many valid forms remain unseen in the training data
  - unseen morphological variants cannot be produced/translated
  - results in bad translation quality

Translating morphologically complex languages

- Morphologically rich languages: large amount of word forms
- Problematic for machine translation:
  - many valid forms remain unseen in the training data
  - unseen morphological variants cannot be produced/translated
  - results in bad translation quality

- SMT systems
  - can only translate and output words observed in the training data
  - cannot handle unseen words

Translating morphologically complex languages

- Morphologically rich languages: large amount of word forms
- Problematic for machine translation:
  - many valid forms remain unseen in the training data
  - unseen morphological variants cannot be produced/translated
  - results in bad translation quality

- SMT systems
  - can only translate and output words observed in the training data
  - cannot handle unseen words

- NMT systems
  - typically some sort of pre-processing to keep vocabulary size manageable, such as frequency-based segmentation
  - to a certain extent, can handle unseen words
  - rich morphology still not optimally represented

Outline

Coverage of inflected forms in the training data

- Many inflectional variants remain unseen in the training data
- Substantial problem in low-resource settings, but still a problem with larger training corpora

Coverage of inflected forms in the training data

- Many inflectional variants remain unseen in the training data
- Substantial problem in low-resource settings, but still a problem with larger training corpora

- Example: morphological forms of the Czech lemma *čéška* (plural of English *kneecap*) in different training data settings

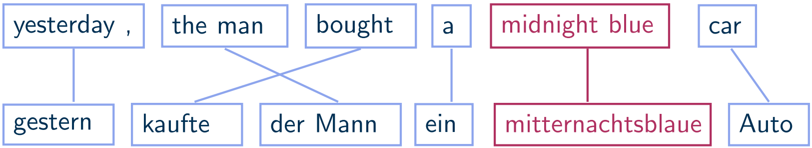
case	surface form	50K	500K	5M	50M
1	čéšky	•	•	•	•
2	čéšek	–	•	•	•
3	čéškám	–	–	•	•
4	čéšky	○	○	•	•
5	čéšky	○	○	○	○
6	čéškách	–	•	•	•
7	čéškami	–	–	–	•

Table 1: 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.

Example taken from Huck et al. (2017)

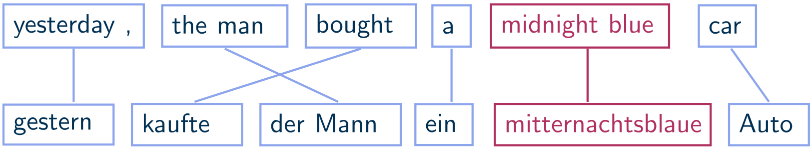
Translating into morphologically rich languages

- Data-sparsity: some inflected forms do not occur in the training data



Translating into morphologically rich languages

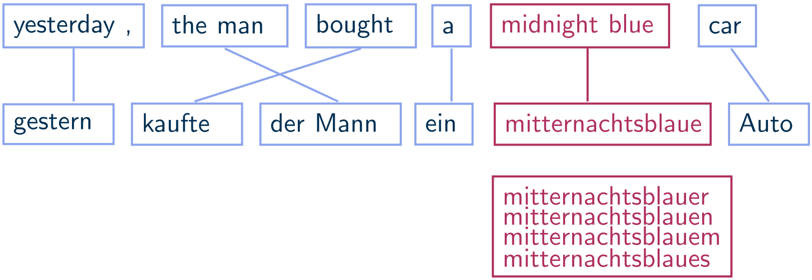
- Data-sparsity: some inflected forms do not occur in the training data



- How can we get the missing inflected forms?

Translating into morphologically rich languages

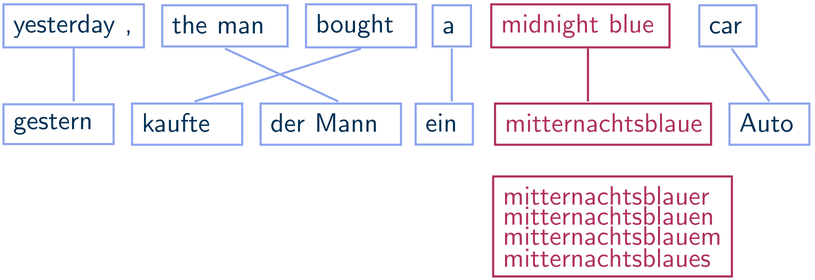
- Data-sparsity: some inflected forms do not occur in the training data



- How can we get the missing inflected forms?
  ⇒ external knowledge resources: e.g. morphological generation tools

Translating into morphologically rich languages

- Data-sparsity: some inflected forms do not occur in the training data



- How can we get the missing inflected forms?

  ⇒ external knowledge resources: e.g. morphological generation tools
- How to select the correct inflected form?

Translating into morphologically rich languages: some strategies

- Increase training data through **back translation**:
  create synthetic parallel data by translating target-side data

  for example Sennrich et al. (2015), Bojar et al. (2011)

Translating into morphologically rich languages: some strategies

- Increase training data through **back translation**:
  create synthetic parallel data by translating target-side data

  for example Sennrich et al. (2015), Bojar et al. (2011)

- Add **synthetic phrases** to the translation phrase table
  to increase coverage of inflected forms (SMT)

  for example Chahuneau et al. (2013), Huck et al. (2017)

Translating into morphologically rich languages: some strategies

- Increase training data through **back translation**:
  create synthetic parallel data by translating target-side data

  for example Sennrich et al. (2015), Bojar et al. (2011)

- Add **synthetic phrases** to the translation phrase table
  to increase coverage of inflected forms (SMT)

  for example Chahuneau et al. (2013), Huck et al. (2017)

- **Two-step approach**: separation of translation and inflection step
  by translating on an abstract representation with subsequent
  generation of inflected forms (SMT+NMT)

  for example Toutanova et al. (2008), Fraser et al. (2012),
  Burlot et al. (2016), Tamchyna et al. (2017)

Outline

Producing unseen morphological variants in SMT (1)

- Idea: generate synthetic morphological variants to add to the phrase-table (for English–Czech translation) Huck et al. (2017)

Producing unseen morphological variants in SMT (1)

- Idea: generate synthetic morphological variants to add to the phrase-table (for English–Czech translation) Huck et al. (2017)

- With a morphological generation tool: synthesize all valid morphological forms from target-side lemmas

Producing unseen morphological variants in SMT (1)

- Idea: generate synthetic morphological variants to add to the phrase-table (for English–Czech translation) Huck et al. (2017)

- With a morphological generation tool: synthesize all valid morphological forms from target-side lemmas
- Newly created morphological variants: add as new translation options

Producing unseen morphological variants in SMT (1)

- Idea: generate synthetic morphological variants to add to the phrase-table (for English–Czech translation) Huck et al. (2017)

- With a morphological generation tool: synthesize all valid morphological forms from target-side lemmas
- Newly created morphological variants: add as new translation options

- Restriction: only use generated variants that fit with the original context (i.e. only some inflectional features can vary, others are kept)

Producing unseen morphological variants in SMT (1)

- Idea: generate synthetic morphological variants to add to the phrase-table (for English–Czech translation) Huck et al. (2017)

- With a morphological generation tool: synthesize all valid morphological forms from target-side lemmas
- Newly created morphological variants: add as new translation options

- Restriction: only use generated variants that fit with the original context (i.e. only some inflectional features can vary, others are kept)

- Scoring the unseen variants: phrase translation and lexical translation probabilities are estimated based on lemmatized forms

Producing unseen morphological variants in SMT (2)

- Training and translation:
  Discriminative classifier that takes into account rich source-side context and dynamically-generated target-side context

Producing unseen morphological variants in SMT (2)

- Training and translation:
  Discriminative classifier that takes into account rich source-side context and dynamically-generated target-side context

- Source-side context: fixed-sized window around the current phrase (with access to lemmas, POS-tags and dependency parses)

Producing unseen morphological variants in SMT (2)

- Training and translation:
  Discriminative classifier that takes into account rich source-side context and dynamically-generated target-side context

- Source-side context: fixed-sized window around the current phrase (with access to lemmas, POS-tags and dependency parses)
- Target-side context: to the left of the current phrase
  → target-side verb-subject agreement
  → agreement within noun phrases/prepositional phrases

Producing unseen morphological variants in SMT (2)

- Training and translation:
  Discriminative classifier that takes into account rich source-side context and dynamically-generated target-side context

- Source-side context: fixed-sized window around the current phrase (with access to lemmas, POS-tags and dependency parses)
- Target-side context: to the left of the current phrase
  → target-side verb-subject agreement
  → agreement within noun phrases/prepositional phrases

- Source-side and target-side features as independent components
  - semantic level: choosing a correct lemma
  - morpho-syntactic level: choosing the correct form
    (tag + morphological features in the given context)

Producing unseen morphological variants in SMT (3)

- Experimental results: substantial improvements in BLEU, in particular for small and medium sized settings

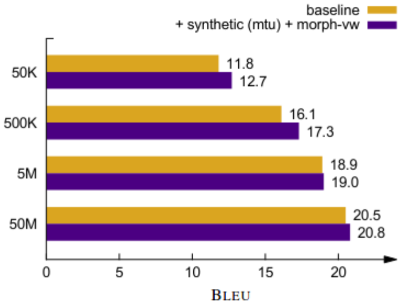


Figure 1: Visualization of the English→Czech translation quality on newstest2016, showing the benefit of our approach under different training resource conditions (50K/500K/5M/50M).

from Huck et al. (2017)

Outline

Two-step inflection generation approach: motivation

Separate **translation step** and **target-side inflection**

Two-step inflection generation approach: motivation

Separate **translation step** and **target-side inflection**

- **Translation** on an abstract target-side representation
  - related inflectional variants are mapped into one form (lemma)
  - inflectional features are kept separately (morph. tag)

    → better generalization

Two-step inflection generation approach: motivation

Separate **translation step** and **target-side inflection**

- **Translation** on an abstract target-side representation
  - related inflectional variants are mapped into one form (lemma)
  - inflectional features are kept separately (morph. tag)
    → better generalization
  - reduce differences between source and target language:
    temporarily remove target-side specific features

Two-step inflection generation approach: motivation

Separate **translation step** and **target-side inflection**

- **Translation** on an abstract target-side representation
  - related inflectional variants are mapped into one form (lemma)
  - inflectional features are kept separately (morph. tag)
    → better generalization
  - reduce differences between source and target language:
    temporarily remove target-side specific features

- **Generation** of target-side inflected forms
  - integration of external knowledge:
    tool for morphological analysis/generation SMOR: Schmid (2005)
  - independent of observed training instances → generate new forms

Two-step inflection generation approach: motivation

Separate **translation step** and **target-side inflection**

- **Translation** on an abstract target-side representation
  - related inflectional variants are mapped into one form (lemma)
  - inflectional features are kept separately (morph. tag)
    → better generalization
  - reduce differences between source and target language:
    temporarily remove target-side specific features

- **Generation** of target-side inflected forms
  - integration of external knowledge:
    tool for morphological analysis/generation SMOR: Schmid (2005)
  - independent of observed training instances → generate new forms

- SMT: nominal inflection Fraser et al. 2012
- NMT: nominal and verbal inflection Tamchyna et al. 2017

SMOR: morphological analysis and generation

- **Analysis**

```
analyze> blaue
------------------------------------------
blau<+ADJ><Pos><Neut><Acc><Sg><Wk>
blau<+ADJ><Pos><Neut><Nom><Sg><Wk>
blau<+ADJ><Pos><Masc><Nom><Sg><Wk>
blau<+ADJ><Pos><NoGend><Acc><Pl><St>
blau<+ADJ><Pos><NoGend><Nom><Pl><St>
blau<+ADJ><Pos><Fem><Acc><Sg><Wk>
blau<+ADJ><Pos><Fem><Acc><Sg><St>
blau<+ADJ><Pos><Fem><Nom><Sg><Wk>
blau<+ADJ><Pos><Fem><Nom><Sg><St>
```

SMOR: morphological analysis and generation

- **Analysis**

```
analyze> blaue
-----------------------------------------
blau<+ADJ><Pos><Neut><Acc><Sg><Wk>
blau<+ADJ><Pos><Neut><Nom><Sg><Wk>
blau<+ADJ><Pos><Masc><Nom><Sg><Wk>
blau<+ADJ><Pos><NoGend><Acc><Pl><St>
blau<+ADJ><Pos><NoGend><Nom><Pl><St>
blau<+ADJ><Pos><Fem><Acc><Sg><Wk>
blau<+ADJ><Pos><Fem><Acc><Sg><St>
blau<+ADJ><Pos><Fem><Nom><Sg><Wk>
blau<+ADJ><Pos><Fem><Nom><Sg><St>
```

- Syncretism: combine with parse analysis to disambiguate in context

SMOR: morphological analysis and generation

- **Analysis**

```
analyze> blaue
-----------------------------------
blau<+ADJ><Pos><Neut><Acc><Sg><Wk>
blau<+ADJ><Pos><Neut><Nom><Sg><Wk>
blau<+ADJ><Pos><Masc><Nom><Sg><Wk>
blau<+ADJ><Pos><NoGend><Acc><Pl><St>
blau<+ADJ><Pos><NoGend><Nom><Pl><St>
blau<+ADJ><Pos><Fem><Acc><Sg><Wk>
blau<+ADJ><Pos><Fem><Acc><Sg><St>
blau<+ADJ><Pos><Fem><Nom><Sg><Wk>
blau<+ADJ><Pos><Fem><Nom><Sg><St>
```

- Syncretism: combine with parse analysis to disambiguate in context

- **Generation**

```
generate> blau<+ADJ><Pos><Fem><Nom><Sg><Wk>
-----------------------------------------------
blaue
```

Inflection prediction approach: data representation

Inflection prediction approach: data representation

Translation model on abstract lemmatized representation

- Inflected forms (nominal phrases) are replaced with lemmas
  blau, *blaue*, *blaues*, *blauem*, *blauen*, *blauer* → `blau<+ADJ><Pos>`
- Some inflectional features are annotated as markup

Inflection prediction approach: data representation

Translation model on abstract lemmatized representation

- Inflected forms (nominal phrases) are replaced with lemmas
  blau, *blaue*, *blaues*, *blauem*, *blauen*, *blauer* → `blau<+ADJ><Pos>`
- Some inflectional features are annotated as markup

Inflect the lemmatized translation output

- Predict inflectional features: *case*, *number*, *gender* and *strong/weak*
- Generation step:

$$\underbrace{\texttt{blau<+ADJ><Pos>}}_{\text{lemma}}\underbrace{\texttt{<Neut><Acc><Sg><Wk>}}_{\text{features}} \rightarrow \texttt{blaue}$$

Features for German Nominal Inflection

	English	**German**
number	✓	target-side number of a phrase is determined by the source-side
gender	∅	innate to the noun
strong/weak inflection	∅	depends on the particular setting of definite/indefinite article, number and case within the NP
case	∅	depends on the syntactic function of the NP (→ semantic dimension)

Features for German Nominal Inflection

	English	German
number	✓	target-side number of a phrase is determined by the source-side
gender	∅	innate to the noun
strong/weak inflection	∅	depends on the particular setting of definite/indefinite article, number and case within the NP
case	∅	depends on the syntactic function of the NP (→ semantic dimension)

Feature prediction:
CRF sequence models trained on local context information

Wapiti toolkit: Lavergne et al. (2010)

Markup for Feature Prediction

- The markup helps to predict inflectional features
- Markup sets values which are propagated over the phrase
  - *add markup* for features that are innate or given by the source-side
  - *no markup* for features that entirely depend on target-side context

Markup for Feature Prediction

- The markup helps to predict inflectional features
- Markup sets values which are propagated over the phrase
  - *add markup* for features that are innate or given by the source-side
  - *no markup* for features that entirely depend on target-side context

			markup
noun	Apfel<+NN><**Masc**><**Sg**>	*apple*	gender, number
adjective	lustig<+ADJ><Pos>	*funny*	∅
article	die<+ART><Def>	*the*	∅
preposition	in<APPR><**Dat**>	*in*	case (positional vs. directional)
verb	kauft<VVFIN>	*buys*	fully inflected

Process of Inflection Prediction

English input … these buses may have access to that country …

SMT output with markup	**predicted features**	**inflected forms**	**gloss**
solche<+INDEF><Pro>			such
Bus<+NN><Masc><Pl>			buses
haben<VAFIN>			have
dann<ADV>			then
zwar<ADV>			though
Zugang<+NN><Masc><Sg>			access
zu<APPR><Dat>			to
die<+ART><Def>			the
betreffend<+ADJ><Pos>			respective
Land<+NN><Neut><Sg>			country

Process of Inflection Prediction

English input … these buses may have access to that country …

SMT output with markup	**predicted features**		**inflected forms**	**gloss**
solche<+INDEF><Pro>	PIAT			such
Bus<+NN><Masc><Pl>	NN-Masc	Pl		buses
haben<VAFIN>	haben<V>			have
dann<ADV>	ADV			then
zwar<ADV>	ADV			though
Zugang<+NN><Masc><Sg>	NN-Masc	Sg		access
zu<APPR><Dat>	APPR-Dat			to
die<+ART><Def>	ART			the
betreffend<+ADJ><Pos>	ADJA			respective
Land<+NN><Neut><Sg>	NN-Neut	Sg		country

Process of Inflection Prediction

English input … these buses may have access to that country …

SMT output with markup	**predicted features**	**inflected forms**	**gloss**
solche<+INDEF><Pro>	PIAT-Masc.Nom.Pl.St		such
Bus<+NN><Masc><Pl>	NN-Masc.Nom.Pl.Wk		buses
haben<VAFIN>	haben<V>		have
dann<ADV>	ADV		then
zwar<ADV>	ADV		though
Zugang<+NN><Masc><Sg>	NN-Masc.Acc.Sg.St		access
zu<APPR><Dat>	APPR-Dat		to
die<+ART><Def>	ART-Neut.Dat.Sg.St		the
betreffend<+ADJ><Pos>	ADJA-Neut.Dat.Sg.Wk		respective
Land<+NN><Neut><Sg>	NN-Neut.Dat.Sg.Wk		country

Process of Inflection Prediction

English input … these buses may have access to that country …

SMT output with markup	**predicted features**	**inflected forms**	**gloss**
solche<+INDEF><Pro>	PIAT-Masc.Nom.Pl.St	**solche**	such
Bus<+NN><Masc><Pl>	NN-Masc.Nom.Pl.Wk	**Busse**	buses
haben<VAFIN>	haben<V>	haben	have
dann<ADV>	ADV	dann	then
zwar<ADV>	ADV	zwar	though
Zugang<+NN><Masc><Sg>	NN-Masc.Acc.Sg.St	**Zugang**	access
zu<APPR><Dat>	APPR-Dat	zu	to
die<+ART><Def>	ART-Neut.Dat.Sg.St	**dem**	the
betreffend<+ADJ><Pos>	ADJA-Neut.Dat.Sg.Wk	**betreffenden**	respective
Land<+NN><Neut><Sg>	NN-Neut.Dat.Sg.Wk	**Land**	country

Results: Inflection Prediction

- English–German phrase-based SMT system (MOSES)
- 4.5M parallel sentences, 5-gram language model of 45M sentences

Results: Inflection Prediction

- English–German phrase-based SMT system (MOSES)
- 4.5M parallel sentences, 5-gram language model of 45M sentences

	tuning 1		tuning 2	
	news'14	**news'15**	**news'14**	**news'15**
Surface	19.17	20.86	19.03	20.80
Inflection Prediction	19.35	21.21*	19.32*	21.16*

*: significant improvement (sample size 1,000 and p-value 0.05)

- Inflection prediction system obtains better results than surface system

Results: Inflection Prediction

- English–German phrase-based SMT system (MOSES)
- 4.5M parallel sentences, 5-gram language model of 45M sentences

	tuning 1		tuning 2	
	news'14	**news'15**	**news'14**	**news'15**
Surface	19.17	20.86	19.03	20.80
Inflection Prediction	19.35	21.21*	19.32*	21.16*

*: significant improvement (sample size 1,000 and p-value 0.05)

- Inflection prediction system obtains better results than surface system
- Similar results in other domains (e.g. medical domain)

Results: Inflection Prediction

- English–German phrase-based SMT system (MOSES)
- 4.5M parallel sentences, 5-gram language model of 45M sentences

	tuning 1		tuning 2	
	news'14	**news'15**	**news'14**	**news'15**
Surface	19.17	20.86	19.03	20.80
Inflection Prediction	19.35	21.21*	19.32*	21.16*

*: significant improvement (sample size 1,000 and p-value 0.05)

- Inflection prediction system obtains better results than surface system
- Similar results in other domains (e.g. medical domain)
- BLEU is not an ideal measure: evidence that BLEU underestimates performance in WMT human evaluation

Example: Inflection Prediction

Input	in particular , the actresses play a major role in the sometimes rather dubious staging .
Surface	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdige Inszenierung .
Inflection Prediction	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdigen Inszenierung .

Example: Inflection Prediction

Input	in particular , the actresses play a major role in the sometimes rather dubious staging .
Surface	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdige Inszenierung .
Inflection Prediction	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdigen Inszenierung .

- Parallel data:
  fragwürdige, fragwürdigen occur with similar frequency,
  no bigram of "*fragwürdig + inszenierung*"

Example: Inflection Prediction

Input	in particular , the actresses play a major role in the sometimes rather dubious staging .
Surface	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdige Inszenierung .
Inflection Prediction	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdigen Inszenierung .

- Parallel data:
  fragwürdige, fragwürdigen occur with similar frequency,
  no bigram of "*fragwürdig + inszenierung*"
- Surface language model: 2 occurrences of *fragwürdige inszenierung*

Example: Inflection Prediction

Input	in particular , the actresses play a major role in the sometimes rather dubious staging .
Surface	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdige Inszenierung .
Inflection Prediction	insbesondere die Schauspielerinnen spielen eine große Rolle in der manchmal etwas fragwürdigen Inszenierung .

- Parallel data:
  fragwürdige, fragwürdigen occur with similar frequency,
  no bigram of "*fragwürdig + inszenierung*"
- Surface language model: 2 occurrences of *fragwürdige inszenierung*
- Stemmed language model representation:
  `fragwürdig[ADJ] Inszenierung<Fem><Sg>[NN]`

Inflection prediction: SMT vs. NMT

- The same approach can also be applied to NMT, with two differences

Inflection prediction: SMT vs. NMT

- The same approach can also be applied to NMT, with two differences
- Modeling of inflectional features:

Inflection prediction: SMT vs. NMT

- The same approach can also be applied to NMT, with two differences

- Modeling of inflectional features:
  - SMT: inflectional features are predicted in a separate model *after* the translation step

Inflection prediction: SMT vs. NMT

- The same approach can also be applied to NMT, with two differences

- Modeling of inflectional features:
  - SMT: inflectional features are predicted in a separate model *after* the translation step
  - NMT: inflectional features are modeled during the translation step
  - NMT systems can handle very long sentences: surface forms can be represented as pairs of lemmas and complex tags (i.e. doubling the sentence length)

Inflection prediction: SMT vs. NMT

- The same approach can also be applied to NMT, with two differences

- Modeling of inflectional features:
  - SMT: inflectional features are predicted in a separate model *after* the translation step
  - NMT: inflectional features are modeled during the translation step
  - NMT systems can handle very long sentences: surface forms can be represented as pairs of lemmas and complex tags (i.e. doubling the sentence length)

- Nominal vs. verbal inflection

Inflection prediction: SMT vs. NMT

- The same approach can also be applied to NMT, with two differences

- Modeling of inflectional features:
  - SMT: inflectional features are predicted in a separate model *after* the translation step
  - NMT: inflectional features are modeled during the translation step
  - NMT systems can handle very long sentences: surface forms can be represented as pairs of lemmas and complex tags (i.e. doubling the sentence length)

- Nominal vs. verbal inflection
  - SMT: only modeling of nominal inflection verbal inflection in this setting is very difficult Ramm et al. (2016)
  - NMT: both nominal and verbal inflection better capturing of global sentence context enables verbal inflection

Inflection prediction: NMT

- Inflection prediction in NMT generally works Tamchyna et al. (2017)
  - English → Czech
  - English → German

Inflection prediction: NMT

- Inflection prediction in NMT generally works — Tamchyna et al. (2017)
  - English → Czech
  - English → German
- What about low-resource scenarios, such as German → Upper Sorbian?
  - very small (parallel) training data set
  - no tool to generate morphology
  - → this language pair is of current interest, but rather difficult ...

Inflection prediction: NMT

- Inflection prediction in NMT generally works — Tamchyna et al. (2017)
  - English → Czech
  - English → German
- What about low-resource scenarios, such as German → Upper Sorbian?
  - very small (parallel) training data set
  - no tool to generate morphology
  - → this language pair is of current interest, but rather difficult ...
- Currently ongoing work: modeling word formation in NMT
  - abstract lemma-tag representation provides a sound basis to integrate further linguistic information
  - learn word-formation processes across languages
  - → later in this talk ...

Outline

Reducing the complexity of words: overview

- Translating *compounds* in SMT
  - Splitting compounds on the source side — Koehn and Knight (2003)
  - Modeling compounds on the target side — Cap et al. (2014)

Reducing the complexity of words: overview

- Translating *compounds* in SMT
  - Splitting compounds on the source side — Koehn and Knight (2003)
  - Modeling compounds on the target side — Cap et al. (2014)

- Handling large vocabulary in NMT
  - Reducing the vocabulary size — Sennrich et al. (2016)
  - Linguistically informed segmentation approaches
    - compound splitting, prefix/suffix splitting — Huck et al. (2017)
    - combining BPE and morphological analysis — Banerjee et al. (2018)
    - modeling word formation — Weller-Di Marco et al. (2020)

Reducing the complexity of words: overview

- Translating *compounds* in SMT
  - Splitting compounds on the source side — Koehn and Knight (2003)
  - Modeling compounds on the target side — Cap et al. (2014)

- Handling large vocabulary in NMT
  - Reducing the vocabulary size — Sennrich et al. (2016)
  - Linguistically informed segmentation approaches
    - compound splitting, prefix/suffix splitting — Huck et al. (2017)
    - combining BPE and morphological analysis — Banerjee et al. (2018)
    - modeling word formation — Weller-Di Marco et al. (2020)

- Compositional representation of complex morphology — Ataman et al. (2018)

Outline

Source-side compound splitting

- Compounding is common in many languages
  (e.g. German, Dutch, Swedish, Finnish, ...)
- Creates an infinite amount of new words that cannot be translated

Source-side compound splitting

- Compounding is common in many languages
  (e.g. German, Dutch, Swedish, Finnish, ...)
- Creates an infinite amount of new words that cannot be translated
- Compounds are built from simpler words
  - those simpler words might occur in the corpus
  - they can then be translated

Source-side compound splitting

- Compounding is common in many languages
  (e.g. German, Dutch, Swedish, Finnish, ...)
- Creates an infinite amount of new words that cannot be translated
- Compounds are built from simpler words
  - those simpler words might occur in the corpus
  - they can then be translated
- Idea: split compound into known components → translate parts

  Koehn et al. (2003)

  - frequency-based compound splitting method that is then refined
  - evaluate source-side compound splitting in English → German translation

Source-side compound splitting

- Compounding is common in many languages
  (e.g. German, Dutch, Swedish, Finnish, ...)
- Creates an infinite amount of new words that cannot be translated
- Compounds are built from simpler words
  - those simpler words might occur in the corpus
  - they can then be translated
- Idea: split compound into known components → translate parts

  Koehn et al. (2003)

  - frequency-based compound splitting method that is then refined
  - evaluate source-side compound splitting in English → German translation
- Transparent vs. semantically opaque compounds
  → we assume that compounds are transparent ...

Compound splitting: getting splitting options

- Enumerate all possible splittings into known words

Compound splitting: getting splitting options

- Enumerate all possible splittings into known words
- Consider fugenelemente (transitional elements or filler letters)
  - insertion/deletion of particular letters between compound components:

    Aktionsplan → *Aktion|Plan*
    Schweigeminute → *schweigen|Minute*

  → removal or addition of known elements (*s*/*es*/*n*)

Compound splitting: getting splitting options

- Enumerate all possible splittings into known words
- Consider fugenelemente (transitional elements or filler letters)
  - insertion/deletion of particular letters between compound components:

    Aktionsplan → *Aktion|Plan*
    Schweigeminute → *schweigen|Minute*

  → removal or addition of known elements (*s/es/n*)

- Splitting options for *Aktionsplan* (*plan for action*)

  aktionsplan
  aktion plan
  aktions plan
  akt ion plan

  → all parts have been observed in the training data

Compound splitting: frequency-based metric

- Splitting metric based on word frequency

Compound splitting: frequency-based metric

- Splitting metric based on word frequency
- Select the split S with the highest geometric mean of word frequencies of its parts p_i (n being the number of parts):

  $$argmax_S \, \Big(\prod_{p_i \in S} count(p_i)\Big)^{\frac{1}{n}}$$

Compound splitting: frequency-based metric

- Splitting metric based on word frequency
- Select the split S with the highest geometric mean of word frequencies of its parts p_i (n being the number of parts):

$$argmax_S \left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}}$$

- *Aktionsplan*

aktionsplan (852)	*actionplan*	852
aktion (960) – plan (710)	*action – plan*	825.6
aktions (5) – plan (710)	*action – plan*	59.6
akt (224) – ion (1) – plan (710)	*act – ion – plan*	54.2

Compound splitting: frequency-based metric

- Splitting metric based on word frequency
- Select the split S with the highest geometric mean of word frequencies of its parts p_i (n being the number of parts):

$$argmax_S \left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}}$$

- *Aktionsplan*

aktionsplan (852)	*actionplan*	852	$\Leftarrow$
aktion (960) – plan (710)	*action – plan*	825.6	
aktions (5) – plan (710)	*action – plan*	59.6	
akt (224) – ion (1) – plan (710)	*act – ion – plan*	54.2	

Compound splitting: frequency-based metric

- Splitting metric based on word frequency
- Select the split S with the highest geometric mean of word frequencies of its parts p_i (n being the number of parts):

  $$argmax_S \left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}}$$

- *Aktionsplan*

aktionsplan (852)	*actionplan*	852	$\Leftarrow$
aktion (960) – plan (710)	*action – plan*	825.6	
aktions (5) – plan (710)	*action – plan*	59.6	
akt (224) – ion (1) – plan (710)	*act – ion – plan*	54.2	

- *Freitag*

frei (885) – tag (1864)	*free – day*	1284.4
freitag (556)	*friday*	556

Compound splitting: frequency-based metric

- Splitting metric based on word frequency
- Select the split S with the highest geometric mean of word frequencies of its parts p_i (n being the number of parts):

$$argmax_S \left(\prod_{p_i \in S} count(p_i) \right)^{\frac{1}{n}}$$

- *Aktionsplan*

aktionsplan (852)	*actionplan*	852	$\Leftarrow$
aktion (960) – plan (710)	*action – plan*	825.6	
aktions (5) – plan (710)	*action – plan*	59.6	
akt (224) – ion (1) – plan (710)	*act – ion – plan*	54.2	

- *Freitag*

frei (885) – tag (1864)	*free – day*	1284.4	$\Leftarrow$
freitag (556)	*friday*	556	

Compound splitting: looking at parallel data

- How are splitting options translated in the English sentence?

Compound splitting: looking at parallel data

- How are splitting options translated in the English sentence?
  - *Aktionsplan* $\rightarrow$ *action plan, plan for action, ...*
  - *Freitag* $\not\rightarrow$ *free day*

Compound splitting: looking at parallel data

- How are splitting options translated in the English sentence?
  - *Aktionsplan* $\rightarrow$ *action plan, plan for action, ...*
  - *Freitag* $\not\rightarrow$ *free day*
- Derive translation lexicon from word-aligned data

Compound splitting: looking at parallel data

- How are splitting options translated in the English sentence?
  - *Aktionsplan → action plan, plan for action, ...*
  - *Freitag ↛ free day*
- Derive translation lexicon from word-aligned data

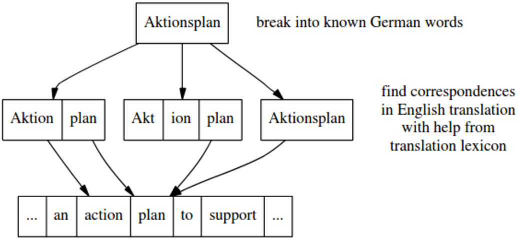


Figure 2: Acquisition of splitting knowledge from a parallel corpus: The split *Aktion–plan* is preferred since it has most coverage with the English (two words overlap)

taken from Koehn et al. (2003)

Compound splitting: looking at parallel data

- How are splitting options translated in the English sentence?
  - *Aktionsplan* → *action plan, plan for action, ...*
  - *Freitag* $\not\rightarrow$ *free day*
- Derive translation lexicon from word-aligned data

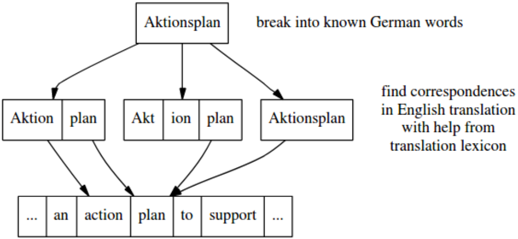


Figure 2: Acquisition of splitting knowledge from a parallel corpus: The split *Aktion–plan* is preferred since it has most coverage with the English (two words overlap)

taken from Koehn et al. (2003)

⇒ Improved splitting precision

Compound splitting: conditioning on Part-Of-Speech

- Only split into content words: nouns, verbs, adjectives, adverbs
  don't split into words such as articles, prepositions or suffixes/prefixes

Compound splitting: conditioning on Part-Of-Speech

- Only split into content words: nouns, verbs, adjectives, adverbs
  don't split into words such as articles, prepositions or suffixes/prefixes

- *folgenden* (*following*) $\not\rightarrow$ *folgen*$_N$ *den*$_{ART}$ (*consequences the*)
  Voraussetzung (*condition*) $\not\rightarrow$ *vor*$_{PREP}$ *aussetzung*$_N$ (PREP *suspension*)

Compound splitting: conditioning on Part-Of-Speech

- Only split into content words: nouns, verbs, adjectives, adverbs
  don't split into words such as articles, prepositions or suffixes/prefixes

- *folgenden* (*following*) $\not\rightarrow$ *folgen*$_N$ *den*$_{ART}$ (*consequences the*)
  Voraussetzung (*condition*) $\not\rightarrow$ *vor*$_{PREP}$ *aussetzung*$_N$ (PREP *suspension*)
  - articles (*der, den, ...*) and *the* are very frequent in the training data
  - similarly: prepositions (*vor, ...*) and its many English translations

Compound splitting: conditioning on Part-Of-Speech

- Only split into content words: nouns, verbs, adjectives, adverbs
  don't split into words such as articles, prepositions or suffixes/prefixes

- *folgenden* (*following*) $\not\to$ $folgen_N$ den_{ART} (*consequences the*)
  Voraussetzung (*condition*) $\not\to$ vor_{PREP} $aussetzung_N$ (PREP *suspension*)
  - articles (*der, den, ...*) and *the* are very frequent in the training data
  - similarly: prepositions (*vor, ...*) and its many English translations

- POS-tag training data, and then obtain word-frequency statistics with POS information

Compound splitting: conditioning on Part-Of-Speech

- Only split into content words: nouns, verbs, adjectives, adverbs
  don't split into words such as articles, prepositions or suffixes/prefixes

- *folgenden* (*following*) $\not\to$ *folgen*$_N$ *den*$_{ART}$ (*consequences the*)
  Voraussetzung (*condition*) $\not\to$ *vor*$_{PREP}$ *aussetzung*$_N$ (PREP *suspension*)
  - articles (*der, den, ...*) and *the* are very frequent in the training data
  - similarly: prepositions (*vor, ...*) and its many English translations

- POS-tag training data, and then obtain word-frequency statistics with POS information

$\Rightarrow$ Improved splitting precision

Source-side compound splitting in SMT: results

- System variants (English → German):

raw	no splits
eager	split into as many parts as possible
freq. based	split into most frequent words
using parallel	using guidance from parallel data
using parallel and POS	as previous, with POS restriction

Source-side compound splitting in SMT: results

- System variants (English → German):

raw	no splits
eager	split into as many parts as possible
freq. based	split into most frequent words
using parallel	using guidance from parallel data
using parallel and POS	as previous, with POS restriction

- Word-based translation

Method	BLEU
raw	0.291
eager	0.222
frequency based	**0.317**
using parallel	0.294
using parallel and POS	0.306

Phrase-based translation

Method	BLEU
raw	0.305
eager	**0.344**
frequency based	**0.342**
using parallel	0.330
using parallel and POS	0.326

taken from Koehn et al. (2003)

Compound splitting in SMT

- So far: compound splitting on the source side
  split compounds: intermediate representation

- How to generate compounds on the target side?
  more difficult: need to generate correctly inflected compounds

Generating target-side compounds in SMT

- How to generate (new) compounds? Cap et al. (2014)

Generating target-side compounds in SMT

- How to generate (new) compounds? Cap et al. (2014)

Pre-processing

- Split compounds into a linguistically informed representation
  - all components look the same throughout the corpus
  - relevant linguistic information is kept

Generating target-side compounds in SMT

- How to generate (new) compounds? Cap et al. (2014)

Pre-processing

- Split compounds into a linguistically informed representation
  - all components look the same throughout the corpus
  - relevant linguistic information is kept

Post-processing

Generating target-side compounds in SMT

- How to generate (new) compounds? Cap et al. (2014)

Pre-processing

- Split compounds into a linguistically informed representation
  - all components look the same throughout the corpus
  - relevant linguistic information is kept

Post-processing

- Merge compounds: `Apfel<NN>` + `Kuchen<NN>` → `Apfelkuchen<NN>`
  - merging decision relies on source-language and target-language features

Generating target-side compounds in SMT

- How to generate (new) compounds? Cap et al. (2014)

Pre-processing

- Split compounds into a linguistically informed representation
  - all components look the same throughout the corpus
  - relevant linguistic information is kept

Post-processing

- Merge compounds: `Apfel<NN>` + `Kuchen<NN>` → `Apfelkuchen<NN>`
  - merging decision relies on source-language and target-language features
- Generation and inflection of compounds
  - generate the correct surface form
  - find the correct inflection (→ combine with inflection-prediction system)

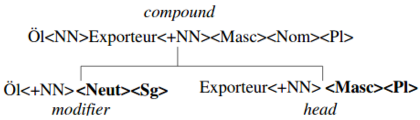
Compound representation

- Linguistically informed compound splitting:
  rule-based morphological analyzer (SMOR) combined with corpus frequencies

Compound representation

- Linguistically informed compound splitting:
  rule-based morphological analyzer (SMOR) combined with corpus frequencies

- Underspecified representation
  reduction to lemmas: keep number+gender information, but remove case



Examples taken from Cap et al. (2014)

Compound representation

- Linguistically informed compound splitting:
  rule-based morphological analyzer (SMOR) combined with corpus frequencies

- Underspecified representation
  reduction to lemmas: keep number+gender information, but remove case

  compound
  Öl<NN>Exporteur<+NN><Masc><Nom><Pl>

  Öl<+NN>**<Neut><Sg>** — Exporteur<+NN> **<Masc><Pl>**
  modifier — *head*

- Representation of modifier and head is the same
  - all components are accessible during training
  - all components can be merged into "new" and "old" compounds

    Haus<+NN><Neut><Sg> + *Boot<+NN><Neut><Pl>*
    → *Haus<NN>Boot<+NN><Neut><Pl>* (merged)

Examples taken from Cap et al. (2014)

Compound merging: when to merge

- How to decide what words to merge? Just merge adjacent nouns?

Compound merging: when to merge

- How to decide what words to merge? Just merge adjacent nouns?
- Merging decision is based on

Compound merging: when to merge

- How to decide what words to merge? Just merge adjacent nouns?

- Merging decision is based on
  - target-side features: various frequencies of words in head position vs. modifier position vs. simplex occurrences

Compound merging: when to merge

- How to decide what words to merge? Just merge adjacent nouns?

- Merging decision is based on
  - target-side features: various frequencies of words in head position vs. modifier position vs. simplex occurrences
  - projected source-side features:
    - English syntactic structure aligned to compound candidate
    - English POS tag
    - alignment features

Compound merging: when to merge

- How to decide what words to merge? Just merge adjacent nouns?

- Merging decision is based on
  - target-side features: various frequencies of words in head position vs. modifier position vs. simplex occurrences
  - projected source-side features:
    - English syntactic structure aligned to compound candidate
    - English POS tag
    - alignment features

merge	ein erhöhtes verkehrs aufkommen sorgt für chaos an increased traffic volume causes chaos (S...(NP(DT an)(VN increased (NN traffic) (NN volume)) ...)
don't merge	für die finanzierung des verkehrs aufkommen pay for the financing of transport (VP(V pay)(PP(IN for)(NP(NP(DT the)(NN financing))(PP(IN of) (NP(NN transport)...))

Compound generation and inflection

- How to recombine components into well-formed compounds?

Compound generation and inflection

- How to recombine components into well-formed compounds?

  take into account transitional elements and "Umlautung"

  - *Ort + Zeit → Ortszeit* (*local time*)
  - *Haus + Fassade → Häuserfassade* (*house front)*

Compound generation and inflection

- How to recombine components into well-formed compounds?

  take into account transitional elements and "Umlautung"
  - *Ort* + *Zeit* → *Ortszeit* (*local time*)
  - *Haus* + *Fassade* → *Häuserfassade* (*house front)*

- Look up combinations of compounds in a list?
  only limited set of compounds

Compound generation and inflection

- How to recombine components into well-formed compounds?

  take into account transitional elements and "Umlautung"

  – *Ort + Zeit → Ortszeit* (*local time*)
  – *Haus + Fassade → Häuserfassade* (*house front*)

- Look up combinations of compounds in a list?
  only limited set of compounds
- Use SMOR to generate compounds
  enables the creation of new compounds

Compound generation and inflection

- How to recombine components into well-formed compounds?

  take into account transitional elements and "Umlautung"
  - *Ort + Zeit → Ortszeit (local time)*
  - *Haus + Fassade → Häuserfassade (house front)*

- Look up combinations of compounds in a list?
  only limited set of compounds
- Use SMOR to generate compounds
  enables the creation of new compounds

- Use inflection prediction system (Fraser et al. 2012)
  to inflect the entire text

Modeling target-side compounds: outcome

- No improvement in BLEU over inflection-prediction baseline
- Manual evaluation showed improved translation of compounds, including the creation of new compounds

Modeling target-side compounds: outcome

- No improvement in BLEU over inflection-prediction baseline
- Manual evaluation showed improved translation of compounds, including the creation of new compounds

- Examples for compound translations

reference	English source	UNSPLIT baseline		STR	
Teddybären	teddy bear	4b	Teddy tragen (Teddy, to bear)	1a	Teddybären (teddy bear)
Emissionsreduktion	emissions reduction	3b	Emissionen Reduzierung (emissions, reducing)	3a	Emissionsverringerung (emission decrease)
Geldstrafe	fine	4b	schönen (fine/nice)	3a	Bußgeld (monetary fine)
Tischtennis	table tennis	2b	Tisch Tennis (table, tennis)	4a	Spieltischtennis (play table tennis)
Kreditkartenmarkt	credit-card market	2b	Kreditkarte Markt (credit-card, market)	4a	Kreditmarkt (credit market)
Rotationstempo	rotation rate	2b	Tempo Rotation (rate, rotation)	4a	Temporotation (rate rotation)

Table 6: Examples of the detailed manual compound analysis for **UNSPLIT** and **STR**.

taken from Cap et al. (2003)

Outline

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation
  - initial vocabulary: character vocabulary

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation
  - initial vocabulary: character vocabulary
  - words are represented as sequence of characters + end-of-word symbol

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation
  - initial vocabulary: character vocabulary
  - words are represented as sequence of characters + end-of-word symbol
  - merge operation: replace the most frequent sequence "a b" → "ab"

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation
  - initial vocabulary: character vocabulary
  - words are represented as sequence of characters + end-of-word symbol
  - merge operation: replace the most frequent sequence "a b" → "ab"
  - continue merging until the desired vocabulary size is reached

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation
  - initial vocabulary: character vocabulary
  - words are represented as sequence of characters + end-of-word symbol
  - merge operation: replace the most frequent sequence "a b" → "ab"
  - continue merging until the desired vocabulary size is reached

- BPE leads to improvements in BLEU and is widely used

Reducing the vocabulary size with BPE

- NMT systems typically operate with a fixed vocabulary
- How to handle open-vocabulary translation?
- Encode rare and unknown words as sequences of sub-word units

- Byte Pair Encoding (BPE) Sennrich et al. 2016
- Simple, frequency-based approach for word segmentation
  - initial vocabulary: character vocabulary
  - words are represented as sequence of characters + end-of-word symbol
  - merge operation: replace the most frequent sequence "a b" → "ab"
  - continue merging until the desired vocabulary size is reached

- BPE leads to improvements in BLEU and is widely used

- Obtained segmentation is often not lingusitically optimal

  Forschungsinstituten (*research institutes*)

  Forschungs|*instituten* vs. *Forsch*|*ungsinstitu*|*ten*

Linguistically informed segmentation

- How can we improve BPE splitting?

Linguistically informed segmentation

- How can we improve BPE splitting?

- Linguistically informed extension of BPE — Huck et al. (2017)
  - Compound splitting
  - Suffix splitting
  - Prefix splitting
  - BPE
  - Cascaded application of the above

Linguistically informed segmentation

- How can we improve BPE splitting?

- Linguistically informed extension of BPE Huck et al. (2017)
  - Compound splitting
  - Suffix splitting
  - Prefix splitting
  - BPE
  - Cascaded application of the above

- Reduction of data sparsity
  - better generalization over morphological variants
  - better lexical selection through compound splitting and separating affixes

Linguistically informed segmentation

- How can we improve BPE splitting?
- Linguistically informed extension of BPE — Huck et al. (2017)
  - Compound splitting
  - Suffix splitting
  - Prefix splitting
  - BPE
  - Cascaded application of the above
- Reduction of data sparsity
  - better generalization over morphological variants
  - better lexical selection through compound splitting and separating affixes
- Better open vocabulary translation
  - generation of new compounds or morphological variants (stem+suffix)
  - better learning of word formation processes
    through linguistic segmentation

Segmentation strategy (1)

- Compound splitting
  - frequency-based compound splitting from Koehn et al. (2003)
  - segment words into parts such that the geometric mean of the parts' frequencies is maximized

Segmentation strategy (1)

- Compound splitting
  - frequency-based compound splitting from Koehn et al. (2003)
  - segment words into parts such that the geometric mean of the parts' frequencies is maximized

- Suffix splitting
  - Split off suffixes with a modified version of the *Porter Stemmer*
  - inflectional suffixes
  - derivational suffixes: nominalization and adjectivization

Segmentation strategy (1)

- Compound splitting
  - frequency-based compound splitting from Koehn et al. (2003)
  - segment words into parts such that the geometric mean of the parts' frequencies is maximized

- Suffix splitting
  - Split off suffixes with a modified version of the *Porter Stemmer*
  - inflectional suffixes
  - derivational suffixes: nominalization and adjectivization

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

Set of suffixes from Huck et al. (2017)

Examples: relations between English and German suffixes

-los with consistent English counterpart *-less*
taktlos – tactless
reglos – motionless
rastlos – restless
schamlos – shameless
German participles ending with *-end*
hängend – hanging
stehend – standing
schlafend – sleeping
lachend – laughing

Examples from Huck et al. (2017)

Segmentation strategy (2)

- Prefix splitting
  - Split off prefixes with a modified version of the *Porter Stemmer*

Segmentation strategy (2)

- Prefix splitting
  - Split off prefixes with a modified version of the *Porter Stemmer*
  - Prefixes tend to change the semantics of the word stem (e.g. negation)

Segmentation strategy (2)

- Prefix splitting
  - Split off prefixes with a modified version of the *Porter Stemmer*
  - Prefixes tend to change the semantics of the word stem (e.g. negation)

prefixes
ab-, an-, anti-, auf-, aus-, auseinander-, außer-, be-, bei-, binnen-, bitter-, blut-, brand-, dar-, des-, dis-, durch-, ein-, empor-, endo-, ent-, entgegen-, entlang-, entzwei-, epi-, er-, extra-, fehl-, fern-, fest-, fort-, frei-, für-, ge-, gegen-, gegenüber-, grund-, heim-, her-, hetero-, hin-, hinter-, hinterher-, hoch-, homo-, homöo-, hyper-, hypo-, inter-, intra-, iso-, kreuz-, los-, miss-, mit-, mono-, multi-, nach-, neben-, nieder-, non-, pan-, para-, peri-, poly-, post-, pro-, prä-, pseudo-, quasi-, schein-, semi-, stock-, sub-, super-, supra-, tief-, tod-, trans-, ultra-, um-, un-, unab-, unan-, unauf-, unaus-, unbe-, unbei-, undar-, undis-, undurch-, unein-, unent-, uner-, unfehl-, unfort-, unfrei-, unge-, unher-, unhin-, unhinter-, unhoch-, unmiss-, unmit-, unnach-, unter-, untief-, unum-, ununter-, unver-, unvor-, unweg-, unwider-, unzer-, unzu-, unüber-, ur-, ver-, voll-, vor-, voran-, voraus-, vorüber-, weg-, weiter-, wider-, wieder-, zer-, zu-, zurecht-, zurück-, zusammen-, zuwider-, über-

Set of prefixes from Huck et al. (2017)

Cascading linguistic segmentation with BPE

- Splitting with BPE allows to reduce the vocabulary to a particular size
- For further reduction of vocabulary:
  apply BPE in addition to previous segmentation approaches

Cascading linguistic segmentation with BPE

- Splitting with BPE allows to reduce the vocabulary to a particular size
- For further reduction of vocabulary:
  apply BPE in addition to previous segmentation approaches

- BPE benefits from the linguistic segmentation
  - inflectional suffixes already split off:
    no more arbitrary splitting of the last characters
  - compound/prefix splitting:
    meaningful sub-word units provide a better basis for BPE splitting

Linguistically informed segmentation in NMT

- Reversibility
  - target-side segmentation needs to be reversible in post-processing: introduce special markup

Linguistically informed segmentation in NMT

- Reversibility
  - target-side segmentation needs to be reversible in post-processing: introduce special markup
  - at the beginning of suffix tokens (`$$`) and the end of prefix tokens (`$$`)
  - between compound parts (`@@`)
    (also important for transitional elements)
  - for upper-casing and lower-casing of word parts (`#U, #L `)

Linguistically informed segmentation in NMT

- Reversibility
  - target-side segmentation needs to be reversible in post-processing: introduce special markup
  - at the beginning of suffix tokens (`$$`) and the end of prefix tokens (`$$`)
  - between compound parts (`@@`)
    (also important for transitional elements)
  - for upper-casing and lower-casing of word parts (`#U, #L`)

Kleinunternehmen	`#U klein Unternehm $$en`	*small enterprise*
irreführende	`#L Irre führ $$end $$e`	*misleading*

Linguistically informed segmentation in NMT

- Reversibility
  - target-side segmentation needs to be reversible in post-processing: introduce special markup
  - at the beginning of suffix tokens (`$$`) and the end of prefix tokens (`$$`)
  - between compound parts (`@@`)
    (also important for transitional elements)
  - for upper-casing and lower-casing of word parts (`#U, #L`)

Kleinunternehmen	`#U klein Unternehm $$en`	*small enterprise*
irreführende	`#L Irre führ $$end $$e`	*misleading*

- Experimental results:
  Improved translation quality with +0,5 BLEU and −0.9 TER
  for English→German translation

Morphologically guided segmentation

- Use a morphological analyzer (e.g. *Morfessor*) to guide segmentation of words into morphs   Banerjee et al. (2018)

Morphologically guided segmentation

- Use a morphological analyzer (e.g. *Morfessor*) to guide segmentation of words into morphs Banerjee et al. (2018)

- Morphological analysis on source side and target side

Morphologically guided segmentation

- Use a morphological analyzer (e.g. *Morfessor*) to guide segmentation of words into morphs Banerjee et al. (2018)
- Morphological analysis on source side and target side
- Comparison of translating lexically close and distant languages
  - English–Hindi, English–Bengali, Bengali–Hindi
  - Linguistically distant language-pairs: Morfessor-based segmentation is better than BPE
  - Linguistically close language-pairs: BPE is better
  - Combined segmentation of Morfessor and BPE is best

Outline

Modeling word formation in NMT

- Lack of generalization in word-level approaches to NMT at the level of inflectional variants and derivations of shared word stems
- Productive word formation: high number of infrequent words

Modeling word formation in NMT

- Lack of generalization in word-level approaches to NMT at the level of inflectional variants and derivations of shared word stems
- Productive word formation: high number of infrequent words
- Linguistically motivated segmentation on source and target side to learn productive word formation processes across languages

  ungovernability $\leftrightarrow$ *Unregierbarkeit*

  un_{PREF} $govern_{V}$ $able_{SUFF-ADJ}$ $ity_{SUFF-NOUN}$

  un_{PREF} $regieren_{V}$ $bar_{SUFF-ADJ}$ $keit_{SUFF-NOUN}$

Modeling word formation in NMT

- Lack of generalization in word-level approaches to NMT at the level of inflectional variants and derivations of shared word stems
- Productive word formation: high number of infrequent words
- Linguistically motivated segmentation on source and target side to learn productive word formation processes across languages

  ungovernability ↔ *Unregierbarkeit*

  un_{PREF} $govern_{V}$ $able_{SUFF-ADJ}$ $ity_{SUFF-NOUN}$

  un_{PREF} $regieren_{V}$ $bar_{SUFF-ADJ}$ $keit_{SUFF-NOUN}$

- Sound morphological processing:
  - better generalization on the word-level and morpheme-level
  - model processes such as compounding and derivation
  - enables the generation of new words

Linguistically sound segmentation

- Frequency-based segmentation approaches (BPE):
  effective, but linguistically uninformed → suboptimal splitting

Linguistically sound segmentation

- Frequency-based segmentation approaches (BPE):
  effective, but linguistically uninformed → suboptimal splitting
- Cannot handle non-concatenative processes
  - umlautung: *Baum*$_{Sg}$ → *Bäume*$_{Pl}$ (*tree/trees*)
  - transitional elements: *Grenz|kontroll|politik* → *Grenze, Kontrolle*
    (*border control policy*)
  - derivation: *abundant* ↔ *abundance*

Linguistically sound segmentation

- Frequency-based segmentation approaches (BPE): effective, but linguistically uninformed → suboptimal splitting
- Cannot handle non-concatenative processes
  - umlautung: $Baum_{Sg} \rightarrow Bäume_{Pl}$ (*tree/trees*)
  - transitional elements: *Grenz|kontroll|politik* → *Grenze, Kontrolle* (*border control policy*)
  - derivation: *abundant* ↔ *abundance*

- Segmentation strategy that takes into account fusional morphology
  - implementing an English morphological analyzer
  - exploiting an existing tool for German

⇒ Obtain a consistent linguistics-informed sub-word representation

Modeling word formation: source-side analysis

- Frequency-based splitting method Koehn et al. (2003)

Modeling word formation: source-side analysis

- Frequency-based splitting method — Koehn et al. (2003)
- Operates on lemmatized data with prefix/suffix information
- Rules for non-concatenative transitions: *beautiful* → $\textit{beauty}_N$ $\textit{ful}_{SUFF}$

Modeling word formation: source-side analysis

- Frequency-based splitting method — Koehn et al. (2003)
- Operates on lemmatized data with prefix/suffix information
- Rules for non-concatenative transitions: $beauti\textit{f}ul \rightarrow beauty_N\ ful_{SUFF}$
- POS information:
  - provides flat word-internal structure
  - guides analysis: $decent_{ADJ} \not\rightarrow de_{PREF}\ cent_N$

Modeling word formation: source-side analysis

- Frequency-based splitting method Koehn et al. (2003)
- Operates on lemmatized data with prefix/suffix information
- Rules for non-concatenative transitions: *beautiful* → $\textit{beauty}_N$ $\textit{ful}_{SUFF}$
- POS information:
  - provides flat word-internal structure
  - guides analysis: $\textit{decent}_{ADJ} \not\rightarrow \textit{de}_{PREF}\ \textit{cent}_N$

word	**analysis**
conspiracy	conspire|V acy|SUFF/N/e
conspiratorial	conspire|V ator|SUFF/N/e ial|SUFF/ADJ/-
conspirator	conspire|V ator|SUFF/N/e
conspire	conspire|V
acquire	acquire|V
acquisition	acquire|V ition|SUFF/N/s→re
acquisitive	acquire|V itive|SUFF/ADJ/s→re
acquisitiveness	acquire|V itive|SUFF/ADJ/s→re ness|SUFF/N/-

Modeling word formation: target-side morphology

- Handle target-side inflection: use lemma-tag generation approach

Modeling word formation: target-side morphology

- Handle target-side inflection: use lemma-tag generation approach
- Selection of lemma analyses
  - the lemma representation is obtained from SMOR
    many analyses at different levels of granularity
  - carefully select lemma representation → basis for further segmentation
  - combine SMOR analyses with word frequencies

Modeling word formation: target-side morphology

- Handle target-side inflection: use lemma-tag generation approach
- Selection of lemma analyses
  - the lemma representation is obtained from SMOR
    many analyses at different levels of granularity
  - carefully select lemma representation → basis for further segmentation
  - combine SMOR analyses with word frequencies

Word	*atomwaffenfrei*
SMOR	Atom<NN>Waffe<NN>frei<+ADJ>
	nuclear weapon free
Word	*Forschungsergebnis*
SMOR	forschen<V>ung<NN><SUFF>Ergebnis<+NN>
	research result
Word	*gefährlich*
SMOR	Gefahr<NN>lich<SUFF><+ADJ>
	danger -ous

Modeling word formation: target-side morphology

- Handle target-side inflection: use lemma-tag generation approach
- Selection of lemma analyses
  - the lemma representation is obtained from SMOR
    many analyses at different levels of granularity
  - carefully select lemma representation → basis for further segmentation
  - combine SMOR analyses with word frequencies

Word	*atomwaffenfrei*
SMOR	`Atom<NN>Waffe<NN>frei<+ADJ>`
	nuclear weapon free
Word	*Forschungsergebnis*
SMOR	`forschen<V>ung<NN><SUFF>Ergebnis<+NN>`
	research result
Word	*gefährlich*
SMOR	`Gefahr<NN>lich<SUFF><+ADJ>`
	danger -ous

Translation experiments: data representation

- Lemma-tag representation on source and target side → generalization

Translation experiments: data representation

- Lemma-tag representation on source and target side → generalization
- Segmentation based on morphological analysis
  (combined with BPE to reach vocabulary size)

Translation experiments: data representation

- Lemma-tag representation on source and target side → generalization
- Segmentation based on morphological analysis
  (combined with BPE to reach vocabulary size)
- **German** compound splitting, splitting of nominalization suffixes

Translation experiments: data representation

- Lemma-tag representation on source and target side → generalization
- Segmentation based on morphological analysis
  (combined with BPE to reach vocabulary size)
- **German** compound splitting, splitting of nominalization suffixes
- **English** variation of markup and splitting granularity

EN Morph-Markup-Split
enthusiasm <N> tic<SUFF_ADJ> ally<SUFF_ADV> explode <V> ion<SUFF_N>

EN Morph-noMarkup-Split
enthusiasm tic<SUFF_ADJ> ally<SUFF_ADV> explode ion<SUFF_N>

EN Morph-noMarkup-noSplit
enthusiasmtic<SUFF_ADJ>ally<SUFF_ADV> explodeion<SUFF_N>

Translation experiments: data representation

- Lemma-tag representation on source and target side → generalization
- Segmentation based on morphological analysis
  (combined with BPE to reach vocabulary size)
- **German** compound splitting, splitting of nominalization suffixes
- **English** variation of markup and splitting granularity

EN Morph-Markup-Split
enthusiasm <N> tic<SUFF_ADJ> ally<SUFF_ADV> explode <V> ion<SUFF_N>

EN Morph-noMarkup-Split
enthusiasm tic<SUFF_ADJ> ally<SUFF_ADV> explode ion<SUFF_N>

EN Morph-noMarkup-noSplit
enthusiasmtic<SUFF_ADJ>ally<SUFF_ADV> explodeion<SUFF_N>

- Non-split morphological analyses: enables BPE splitting
  into valid and existing sub-words

Translation experiments: results

- English → German NMT transformer model on news data
- Compare different training data settings: 250k ↔ 4M sentences

Translation experiments: results

- English → German NMT transformer model on news data
- Compare different training data settings: 250k ↔ 4M sentences

	Source (EN)	Target (DE)	Small (250k)	Medium (1M)	Large (2M)	Larger (4M)
1	plain	plain	21.77	26.60	28.66	33.71
2	plain	oldLemTag	22.25	26.96	28.87	33.97
3	plain	LemTag	22.47	27.05	28.61	33.90
4	LemTag	LemTag	23.32	27.36	28.88	34.28
5	LemTag	LemTagSplit	22.55	27.22	29.07	34.21
6	LemTag Markup-Split	LemTag	21.85	26.90	29.33	33.96
7	LemTag noMarkup-Split	LemTag	22.86	27.05	29.20	34.10
8	LemTag noMarkup-noSplit	LemTag	22.82	27.18	29.18	34.12
9	LemTag Markup-Split	LemTagSplit	22.25	27.12	29.39	34.38
10	LemTag noMarkup-Split	LemTagSplit	22.53	26.90	29.10	34.12
11	LemTag noMarkup-noSplit	LemTagSplit	23.23	27.55	29.42	34.19

- Improvements over standard lemma-tag system
- Best system variants: morphological analysis on source and target side

Translation experiments: example

- Example from the medical domain (Large 4M setting):

Input	normally involves coagulation tests on the patient's blood.
Surface	beinhalten normalerweise Coagulationstests am Blut des Patienten.
Morph	handelt es sich in der Regel um Gerinnungstests am Blut des Patienten.
Ref	beinhaltet normalerweise Gerinnungstests der Blut des Patienten.

Translation experiments: example

- Example from the medical domain (Large 4M setting):

Input	normally involves coagulation tests on the patient's blood.
Surface	beinhalten normalerweise Coagulationstests am Blut des Patienten.
Morph	handelt es sich in der Regel um Gerinnungstests am Blut des Patienten.
Ref	beinhaltet normalerweise Gerinnungstests der Blut des Patienten.

- Segmentation of *coagulation* (f=19) and *coagulate* (f=3)

Surface (BPE)	Morph. System (morph + BPE)
co@@ ag@@ ulation	co@@ ag@@ ulate ion<SUFF_N>
co@@ ag@@ ulate	co@@ ag@@ ulate

  - even with BPE splitting: better generalization
  - enables matches between *coagulate* and *coagulation* (and other inflected variants: *coagulates*, *coagulated*, ...)

Recap: inflection and segmentation strategies

- Methods to model inflectional morphology for SMT and NMT
  - generally successful ...
  - inflection-prediction: basis for target-side linguistic modeling
  - target-side compound generation, modeling word formation
  - modeling complement types: subcategorization and choice of prepositions Weller-Di Marco et al. (2016)

Recap: inflection and segmentation strategies

- Methods to model inflectional morphology for SMT and NMT
  - generally successful ...
  - inflection-prediction: basis for target-side linguistic modeling
  - target-side compound generation, modeling word formation
  - modeling complement types: subcategorization and choice of prepositions Weller-Di Marco et al. (2016)
- Segmentation approaches with varying degrees of complexity
  - generally successful ...
  - modeling word formation: currently ongoing research

Recap: inflection and segmentation strategies

- Methods to model inflectional morphology for SMT and NMT
  - generally successful ...
  - inflection-prediction: basis for target-side linguistic modeling
  - target-side compound generation, modeling word formation
  - modeling complement types: subcategorization and choice of prepositions Weller-Di Marco et al. (2016)
- Segmentation approaches with varying degrees of complexity
  - generally successful ...
  - modeling word formation: currently ongoing research
- Potential problems
  - at least to some extent language-specific
  - resource-intensive: requires morphological annotation tools, parsers, ...
  - morphological analysis is error prone → errors in translation

Outline

Compositional representation of morphologically-rich input

- Replace source-language embeddings with a bi-directional RNN that generates compositional representations of the input

Ataman et al. (2018)

Compositional representation of morphologically-rich input

- Replace source-language embeddings with a bi-directional RNN that generates compositional representations of the input

Ataman et al. (2018)

- Obtain input representation from composing smaller units, such as character n-grams
- Composition to learn morphology and lexical meaning in a bilingual context
- Composition layer computes final input representation passed to the encoder to generate translations

Compositional representation of morphologically-rich input

- Replace source-language embeddings with a bi-directional RNN that generates compositional representations of the input

  Ataman et al. (2018)

- Obtain input representation from composing smaller units, such as character n-grams
- Composition to learn morphology and lexical meaning in a bilingual context
- Composition layer computes final input representation passed to the encoder to generate translations
- Avoids explicit and potentially sub-optimal segmentation

Compositional representation of morphologically-rich input

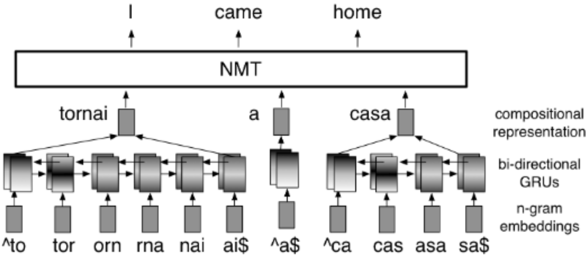


Figure 1: Translation of the Italian sentence *tor-nai a casa* (*I came home*) with a word-level representation composed from character trigrams.

Taken from Ataman et al. (2018)

Languages and Results

- Experiments with five languages from different morphological typologies in a low-resource setting (translating into English)

Language	Morphological Typology	Morphological Complexity
Turkish	*Agglutinative*	*High*
Arabic	*Templatic*	*High*
Czech	*Fusional, Agglutinative*	*High*
German	*Fusional*	*Medium*
Italian	*Fusional*	*Low*

Table 1: The languages evaluated in our study and their morphological characteristics.

Language Pair	# tokens Src	# tokens Tgt	# types Src	# types Tgt
Tr - En	2,7M	2,0M	171K	53K
Ar - En	3,9M	4,9M	220K	120K
Cs - En	2,0M	2,3M	118K	50K
De - En	4,0M	4,3M	144K	69K
It - En	3,5M	3,8M	95K	63K

Table 2: Sizes of the training sets and vocabularies in the TED Talks benchmark. Development and test sets are on average 50K to 100K tokens. (*M*: Million, *K*: Thousand.)

Taken from Ataman et al. (2018)

Languages and Results

- Experiments with five languages from different morphological typologies in a low-resource setting (translating into English)

Language	Morphological Typology	Morphological Complexity
Turkish	*Agglutinative*	*High*
Arabic	*Templatic*	*High*
Czech	*Fusional, Agglutinative*	*High*
German	*Fusional*	*Medium*
Italian	*Fusional*	*Low*

Table 1: The languages evaluated in our study and their morphological characteristics.

Language Pair	# tokens Src	# tokens Tgt	# types Src	# types Tgt
Tr - En	2,7M	2,0M	171K	53K
Ar - En	3,9M	4,9M	220K	120K
Cs - En	2,0M	2,3M	118K	50K
De - En	4,0M	4,3M	144K	69K
It - En	3,5M	3,8M	95K	63K

Table 2: Sizes of the training sets and vocabularies in the TED Talks benchmark. Development and test sets are on average 50K to 100K tokens. (*M*: Million, *K*: Thousand.)

Taken from Ataman et al. (2018)

- Results: compositional models improve over simple BPE models
- Best setting: character trigrams as input symbols and words as final input representation

Outline

Modeling syntax in machine translation

- Different syntactic structures are hard to capture in machine translation

Modeling syntax in machine translation

- Different syntactic structures are hard to capture in machine translation
- SMT: long distance-reordering is costly and sometimes impossible
  NMT: can capture long-distance relations, but can still benefit from syntactic information

Modeling syntax in machine translation

- Different syntactic structures are hard to capture in machine translation
- SMT: long distance-reordering is costly and sometimes impossible
  NMT: can capture long-distance relations, but can still benefit from syntactic information

- SMT: reordering as pre-processing — Collins et al. (2005)

Modeling syntax in machine translation

- Different syntactic structures are hard to capture in machine translation
- SMT: long distance-reordering is costly and sometimes impossible
  NMT: can capture long-distance relations, but can still benefit from syntactic information
- SMT: reordering as pre-processing Collins et al. (2005)
- Syntactic information in NMT
  - Modeling target syntax through CCG tags Nadejde et al. (2017)
  - More strategies to model Syntax in NMT

Source-side reordering in SMT

- Different syntactic structures are hard to capture in word alignment
  for example: placement of verbs in English and German

Source-side reordering in SMT

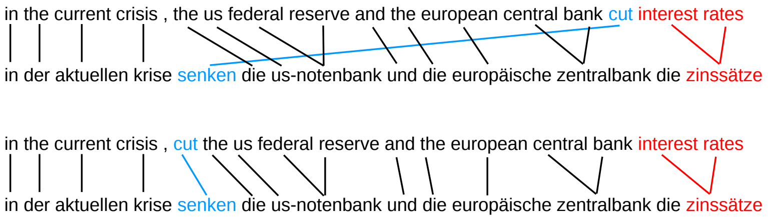
- Different syntactic structures are hard to capture in word alignment
  for example: placement of verbs in English and German
- Pre-processing step:
  reorder source-side such that it adopts the target-side structure

Colins et al. (2005)

Source-side reordering in SMT

- Different syntactic structures are hard to capture in word alignment
  for example: placement of verbs in English and German
- Pre-processing step:
  reorder source-side such that it adopts the target-side structure

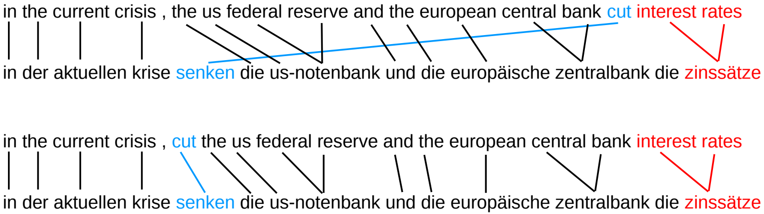
Colins et al. (2005)



Source-side reordering in SMT

- Different syntactic structures are hard to capture in word alignment
  for example: placement of verbs in English and German
- Pre-processing step:
  reorder source-side such that it adopts the target-side structure

Colins et al. (2005)



- Source-side reordering typically leads to improvements in BLEU

Modeling target syntax through CCG tags

- NMT models can partially learn syntactic information
- Some complex syntactic phenomena are poorly modeled

Modeling target syntax through CCG tags

- NMT models can partially learn syntactic information
- Some complex syntactic phenomena are poorly modeled

- Tight integration of words and syntactic information
- Interleaving words with CCG supertags Nadejde et al. (2017)
  - sequences of CCG-tag word pairs
  - added to target-side and source-side (if available for the language pair)

Modeling target syntax through CCG tags

- NMT models can partially learn syntactic information
- Some complex syntactic phenomena are poorly modeled

- Tight integration of words and syntactic information
- Interleaving words with CCG supertags Nadejde et al. (2017)
  - sequences of CCG-tag word pairs
  - added to target-side and source-side (if available for the language pair)

- CCG tags provide global syntactic information
  - subcategorization information
  - attachment
  - tense/morphological aspects of a word in its context

Interleaving with CCG tags: example

Source-side

BPE:	Obama	receives	Net+	an+	yahu	in	the	capital	of	USA
IOB:	O	O	B	I	E	O	O	O	O	O
CCG:	NP	((S[dcl]\NP)/PP)/NP	NP	NP	NP	PP/NP	NP/N	N	(NP\NP)/NP	NP

Target-side

NP Obama ((S[dcl]\NP)/PP)/NP receives NP Net+ an+ yahu PP/NP in NP/N the N capital (NP\NP)/NP of NP USA

Figure 1: Source and target representation of syntactic information in syntax-aware NMT.

Interleaving with CCG tags: example

Source-side

BPE:	Obama	receives	Net+	an+	yahu	in	the	capital	of	USA
IOB:	O	O	B	I	E	O	O	O	O	O
CCG:	NP	((S[dcl]\NP)/PP)/NP	NP	NP	NP	PP/NP	NP/N	N	(NP\NP)/NP	NP

Target-side

NP Obama ((S[dcl]\NP)/PP)/NP receives NP Net+ an+ yahu PP/NP in NP/N the N capital (NP\NP)/NP of NP USA

Figure 1: Source and target representation of syntactic information in syntax-aware NMT.

- There are two PPs with different attachment possibilities
  - *in* → *Netanyahu* or *receives*?
  - *of* → *capital* or *Netanhayu* or *receives*?

Interleaving with CCG tags: example

Source-side										
BPE:	Obama	receives	Net+	an+	yahu	in	the	capital	of	USA
IOB:	O	O	B	I	E	O	O	O	O	O
CCG:	NP	((S[dcl]\NP)/PP)/NP	NP	NP	NP	PP/NP	NP/N	N	(NP\NP)/NP	NP
Target-side										
NP Obama ((S[dcl]\NP)/PP)/NP receives NP Net+ an+ yahu PP/NP in NP/N the N capital (NP\NP)/NP of NP USA										

Figure 1: Source and target representation of syntactic information in syntax-aware NMT.

- There are two PPs with different attachment possibilities
  - *in* → *Netanyahu* or *receives*?
  - *of* → *capital* or *Netanhayu* or *receives*?
- Disambiguation through suptertags
  - *((S[dcl]\NP)/PP)/NP* of *receives* indicates that *in* attaches to the verb
  - *(NP\NP)/NP* of *of* indicates that it attaches to *capital*

Interleaving with CCG tags: results

- German → English and Romanian → English
  - improvement for both language pairs with target-side CCG annotation
  - no CCG tags available for DE/RO:
    additional source-side annotation with dependency labels:

    small improvement for German → English
    more improvement for Romanian → English

Interleaving with CCG tags: results

- German → English and Romanian → English
  - improvement for both language pairs with target-side CCG annotation
  - no CCG tags available for DE/RO:
    additional source-side annotation with dependency labels:

    small improvement for German → English
    more improvement for Romanian → English
- English → German and English → Romanian
  - improvement for both language pairs with source-side CCG annotation

Interleaving with CCG tags: results

- German → English and Romanian → English
  - improvement for both language pairs with target-side CCG annotation
  - no CCG tags available for DE/RO:
    additional source-side annotation with dependency labels:

    small improvement for German → English
    more improvement for Romanian → English

- English → German and English → Romanian
  - improvement for both language pairs with source-side CCG annotation

- Observation: large improvements for longer sentences involving syntactic phenomena such as subordinated clauses and PP attachment

More strategies to model syntax in NMT

More strategies to model syntax in NMT

- String-to-Tree translation Aharoni et al. (2017)
  - propose translation into a serialized constituency tree
  - mixed results for German → English translation using large data
  - consistent improvement for a low-resource setting
    for DE-EN, RU-EN, CS-EN

More strategies to model syntax in NMT

- String-to-Tree translation Aharoni et al. (2017)
  - propose translation into a serialized constituency tree
  - mixed results for German → English translation using large data
  - consistent improvement for a low-resource setting for DE-EN, RU-EN, CS-EN
- Tree-to-Sequence Attentional Neural Machine Translation Eriguchi et al. (2016)
  - propose using a parse tree on the source side to guide the attention model
  - improvements for English–Japanese translation

More strategies to model syntax in NMT

- Graph Convolutional Encoders for Syntax-aware NMT

  Bastings et al. (2017)

  - GCNs use source-side syntactic dependency trees to produce representations of words
  - improvements for English-German and English-Czech

More strategies to model syntax in NMT

- Graph Convolutional Encoders for Syntax-aware NMT

  Bastings et al. (2017)

  - GCNs use source-side syntactic dependency trees to produce representations of words
  - improvements for English-German and English-Czech

- Incorporating Source Syntax into Transformer-Based NMT

  Currey et al. (2019)

  - propose to incorporate constituency parse information
  - leverage linearized parses of the source training sentences
  - multi-task model with a shared encoder/decoder: translate and parse
  - translating from English into 20 languages in low-resource settings: consistent improvements using the multi-task setup
  - no improvements for large-scale settings

Outline

Summary

- Approaches to integrate linguistic information into machine translation
  - phrase-based statistical machine translation
  - neural machine translation

Summary

- Approaches to integrate linguistic information into machine translation
  - phrase-based statistical machine translation
  - neural machine translation

- Focus on modeling morphology
  (inflection, compounding, word formation)
- Brief look into incorporating syntactic information

Summary

- Approaches to integrate linguistic information into machine translation
  - phrase-based statistical machine translation
  - neural machine translation

- Focus on modeling morphology (inflection, compounding, word formation)
- Brief look into incorporating syntactic information

- Integrating linguistic information can lead to improvements on many levels

Thank you!

References

- Roee Aharoni, Yoav Goldberg: *Towards String-to-Tree Neural Machine Translation*. ACL 2017.
- Duygu Ataman, Matteo Negri, Marco Turchi, Marcello Federico: *Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English.* EAMT 2017.
- Duygu Ataman, Marcello Federico: *Compositional Representation of Morphologically-Rich Input for Neural Machine Translation.* ACL 2018.
- Tamali Banerjee, Pushpak Bhattacharya: *Meaningless yet Meaningful: Morphology Grounded Subword-level NMT.* 2nd Workshop on Subword/Character LEvel Models @ ACL 2018.
- Ondřej Bojar, Aleš Tamchyna: *Improving Translation Model by Monolingual Data.* WMT 2011.
- Fabienne Cap, Alexander Fraser, Marion Weller, Aoife Cahill: *How to Produce Unseen Teddy Bears:Improved Morphological Processing of Compounds in SMT*. EACL 2014.
- Michael Collins, Philipp Koehn, and Ivona Kucerova: *Clause Restructuring for Statistical Machine Translation.* ACL 2005.

References

- Alexander Fraser, Marion Weller, Aoife Cahill, Fabienne Cap: *Modeling Inflection and Word-Formation in SMT.* EACL 2012.
- Matthias Huck, Aleš Tamchyna, Ondřej Bojar, Alexander Fraser: *Producing Unseen Morphological Variants in Statistical Machine Translation.* EACL 2017.
- Thomas Lavergne, Olivier Cappé, and Francois Yvon: *Practical very large scale CRFs.* ACL 2010.
- Matthias Huck, Simon Riess, Alexander Fraser: *Target-side Word Segmentation Strategies for Neural Machine Translation.* WMT 2017.
- Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, Alexandra Birch: *Predicting Target Language CCG Supertags Improves Neural Machine Translation*. WMT 2017.
- Anita Ramm, Alexander Fraser: *Modeling verbal inflection for English to German SMT.* WMT 2016.
- Helmut Schmid, Arne Fitschen, and Ulrich Heid: *SMOR: a German Computational Morphology Covering Derivation, Composition, and Inflection.* LREC 2004.
- Rico Sennrich, Barry Haddow, Alexandra Birch: *Improving Neural Machine Translation Models with Monolingual Data.* ACL 2016.

References

- Rico Sennrich, Barry Haddow, Alexandra Birch: *Neural Machine Translation of Rare Words with Subword Units.* ACL 2017.
- Aleš Tamchyna, Marion Weller-Di Marco, Alexander Fraser: *Modeling Target-Side Inflection in Neural Machine Translation.* WMT 2017.
- Kristina Toutanova, Hisami Suzuki, Achim Ruopp: *Applying Morphology Generation Models to Machine Translation.* ACL 2008.
- Marion Weller-Di Marco, Alexander Fraser, Sabine Schulte im Walde: *Modeling Complement Types in Phrase-Based SMT*. WMT 2016.
- Marion Weller-Di Marco, Alexander Fraser: *Modeling Word Formation in English–German Neural Machine Translation*. ACL 2020.