How to account for MWEs in SMT

Jörg Tiedemann

Fabienne Cap

UPPSALA
UNIVERSITY
SWEDEN

Holy what?

Spoiler Alert! This lecture is **not** the holy grail!!
Instead, it is a selection of things you could do...
there is many more!!

Motivation

MWEs that are **frequent** and **continous** are usually not problematic for phrase-based SMT systems

→ They are learned as a phrase in the phrase table

English	German	*Gloss*
by and large	im Großen und Ganzen	*in the great and whole*
begins	fängt an	*catches to*
flea market	Flohmarkt	*flea|market*

Motivation

But: we know better, don't we?

- By far not all MWEs are frequent **and** continous
- In fact, many MWEs cause problems in SMT

Motivation

FACT : MWEs consist of multiple lexical units

: MWEs lead to alignment assymmetries, e.g.

	English	German	*Gloss*
1:1	tree	Baum	*tree*
1:n	begins	fängt an	*catches to*
n:1	apple tree	Apfelbaum	*apple│tree*
n:m	for very little money	für einen Apfel und ein Ei	*for an apple and an egg*

SOLUTION: e.g. compound splitting or particle verb merging

Motivation

FACT : MWEs are rare: many types, not many tokens

: less occurrences → less reliable translations

SOLUTION: enhance frequencies in the training data
→ e.g. through lemmatisation, compound splitting

Motivation

FACT : MWEs are (semantically) non-compositional

: translation of the parts ≠ translation of the whole

EXAMPLE:	kick the ball	=	kick den Ball
but:	kick the bucket	≠	kick den Eimer
		=	sterben (*to die*)

SOLUTION: Tell SMT where MWEs are

Motivation

FACT: MWEs may be discontinous

🇬🇧💻🇩🇪: **impossible?** to learn when phrase size is exceeded

Dazu **leistet** die Effizienz des Vermittlungsverfahrens einen substanziellen **Beitrag**

*To that **make** the effectiveness of the codecision procedure a substantial **contribution**.*

The effectiveness of the codecision procedure has **made** a substantial **contribution** in this case.

SOLUTION: Tell SMT where MWEs are

Motivation

Some new stuff, but the still the same old story....

1. How do we identify MWEs?
2. What do we do once we've found them?

Overview

How to use translations for MWE identification

How to use identified MWEs to improve translation

German:	*da beißt sich die Katze in den Schwanz*
Gloss:	there bites itself the cat into the tail
Literal:	the cat bites itself into the tail
English:	chasing one's tail
	→ going round in circlesl

How to use translations for MWE identification

	MWE characteristics	**Useful?**
1	multiple lexical units	✓
2	many types, not many tokens	?
3	**semantically non compositional**	✓
4	discontinuous components	✗

Procedure

We apply a two-step procedure:

1. extract MWE candidates:
   - focus on verb+PP triples
   - extract from a parallel corpus
   - use a dependency parser
2. identify idiomatic MWEs from these candidates
   - translations are approximated using word alignments
   - **proportion of literal translations**
   - **translational entropy**:
     variance of different translations found

Extract MWE candidates

Extract all verb+PP triples from the German section of Europarl...

Extract MWE candidates

valid	prep	noun	verb	frequency
+	zu	Ausdruck	bringen	4995
	to	*expression*	*bring*	
+	von	Bedeutung	sein	4962
	of	*meaning*	*be*	
+	zu	Kenntnis	nehmen	2740
	to	*knowledge*	*take*	
-	um	Uhr	stattfinden	2725
	at	*clock*	*take place*	
-	nach	Tagesordnung	folgen	2586
	after	*agenda*	*follow*	
+	zu	Verfügung	stehen	2042
	to	*disposal*	*stand*	
-	für	Bericht	stimmen	1812
	for	*report*	*vote*	
+	zu	Verfgüng	stellen	1784
	to	*disposal*	*put*	
+	in	Frage	stellen	1739
	into	*question*	*put*	
-	für	Arbeit	danken	1687
	for	*work*	*thank*	

Using alignments to approximate translations

Context-independent:
("literal translations")

an	=	NO_LINK (72845), **on** (17593), **to** (15268), **in** (13961)
ball	=	**ball** (33), ball court (23), NO_LINK (11), court (3)
macht	=	**power** (2139), NO_LINK (356), can (131), force (64)
bleiben	=	NO_LINK (3309), **remain** (1776), **stay** (290), still (222)

Contex-dependent:
(only when the words occurred in this verb+pp construction)

an	=	**in** (18), NO_LINK (15), **to** (2), **on** (2), follow (1)
macht	=	**power** (28), NO_LINK (3)
bleiben	=	**remain** (11), **stay** (9), NO_LINK (9), retain (2)

an	=	NO_LINK (7), **on** (2)
ball	=	NO_LINK (3), hold line (1), pot boil (1), finger pulse (1), high profile (1), finger (1), **ball** (1)
bleiben	=	NO_LINK (5), keep (2), stick (1), **stay** (1)

Calculate proportion of literal translations

Context-independent:
("literal translations")

an	=	NO_LINK (72845), **on** (17593), **to** (15268), **in** (13961)
ball	=	**ball** (33), ball court (23), NO_LINK (11), court (3)
macht	=	**power** (2139), NO_LINK (356), can (131), force (64)
bleiben	=	NO_LINK (3309), **remain** (1776), **stay** (290), still (222)

Contex-dependent:
(only when the words occurred in this verb+pp construction)

an	=	**in** (18), NO_LINK (15), **to** (2), **on** (2), follow (1)
macht	=	**power** (28), NO_LINK (3)
bleiben	=	**remain** (11), **stay** (9), NO_LINK (9), retain (2)

$\text{an} = \text{in } \frac{18}{23} + \text{to } \frac{2}{23} + \text{on } \frac{2}{23} = \frac{22}{23} = 95\%$
$\text{macht} = 100\%$
$\text{bleiben} = 90\%$
$\text{total} = 95\%$

Calculate proportion of literal translations

Context-independent:
("literal translations")

an	=	NO_LINK (72845), **on** (17593), **to** (15268), **in** (13961)
ball	=	**ball** (33), ball court (23), NO_LINK (11), court (3)
macht	=	**power** (2139), NO_LINK (356), can (131), force (64)
bleiben	=	NO_LINK (3309), **remain** (1776), **stay** (290), still (222)

Contex-dependent:
(only when the words occurred in this verb+pp construction)

an	=	NO_LINK (7), **on** (2)
ball	=	NO_LINK (3), hold line (1), pot boil (1), finger pulse (1), high profile (1), finger (1), **ball** (1)
bleiben	=	NO_LINK (5), keep (2), stick (1), **stay** (1)

an = 95%	an = 100%
macht = 100%	ball = 16%
bleiben = 90%	bleiben = 25%
total = **95%**	total = **45%**

Example Explanation

compositional:
an Macht bleiben = to stay in power

non-compositional:
an Ball bleiben = to hold on

but: may be used compositionally, too!

Calculating translational variance

Context-independent:
("literal translations")

an	=	NO_LINK (72845), on (17593), to (15268), in (13961)
ball	=	ball (33), ball court (23), NO_LINK (11), court (3)
macht	=	power (2139), NO_LINK (356), can (131), force (64)
bleiben	=	NO_LINK (3309), remain (1776), stay (290), still (222)

Contex-dependent:
(only when the words occurred in this verb+pp construction)

4	an	=	in (18), NO_LINK (15), to (2), on (2), follow (1)
1	macht	=	power (28), NO_LINK (3)
3	bleiben	=	remain (11), stay (9), NO_LINK (9), retain (2)

an = in ($\frac{18}{23} ln \frac{18}{23}$) + to ($\frac{2}{23} ln \frac{2}{23}$) + on ($\frac{2}{23} ln \frac{2}{23}$) + follow ($\frac{1}{23} ln \frac{1}{23}$)

macht = 0

bleiben $\approx$ 0,91

total $\approx$ 0,48

Calculating translational variance

Context-independent:
("literal translations")

an	=	NO_LINK (72845), on (17593), to (15268), in (13961)
ball	=	ball (33), ball court (23), NO_LINK (11), court (3)
macht	=	power (2139), NO_LINK (356), can (131), force (64)
bleiben	=	NO_LINK (3309), remain (1776), stay (290), still (222)

Contex-dependent:
(only when the words occurred in this verb+pp construction)

1	an	=	NO_LINK (7), on (2)
6	ball	=	NO_LINK (3), hold line (1), pot boil (1), finger pulse (1), high profile (1), finger (1), ball (1)
3	bleiben	=	NO_LINK (5), keep (2), stick (1), stay (1)

an ≈ 0,53	an = 1
macht = 0	ball ≈ 1.79
bleiben ≈ 0,91	bleiben ≈ 1,03
total ≈ 0,48	total ≈ 0,94

Results

Procedure:

- Extract the 200 most frequent verb+pp triples, then use the two alignment scores to rank them in decreasing order of idiomaticity.
- Evaluate using the uninterpolated average precision (uap), which measures the ranking quality

	uap
baseline	0.587
proportion of literal translations	0.755
translational variance	0.784
adding more restrictions and combinations	0.969

Related Work

Extraction of

- Idiomatic verb+PPs using word alignments (Villada Moiron and Tiedemann, 2006; Fritzinger, 2008)
- English and Portuguese MWEs using 1:n and n:1 word alignments (de Medeiros Casel et al, 2009)
- Hebrew MWEs using a lexicon to generate word for word translations (Tsvetkov and Wintner, 2011)
- ...

Overview

How to use translations for MWE identification

How to use identified MWEs to improve translation

How to use identified MWEs to improve translation

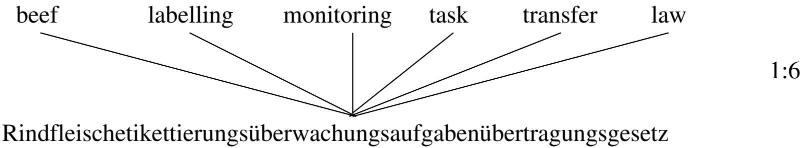
	MWE characteristics	**Solution?**
1	multiple lexical units	**Compound processing**
2	many types, not many tokens	**Lemmatisation**
3	semantically non compositional	Tell SMT where MWEs are
4	discontinuous components	Tell SMT where MWEs are

Compound Splitting for SMT

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

This is a real example!

Compound Splitting for SMT



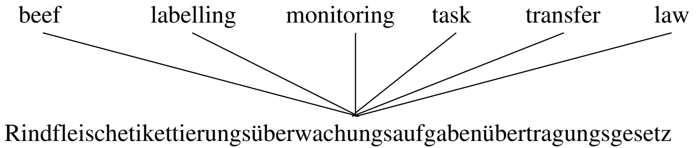
beef labelling monitoring task transfer law

1:6

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

This is a real example!

Compound Splitting for SMT



beef labelling monitoring task transfer law

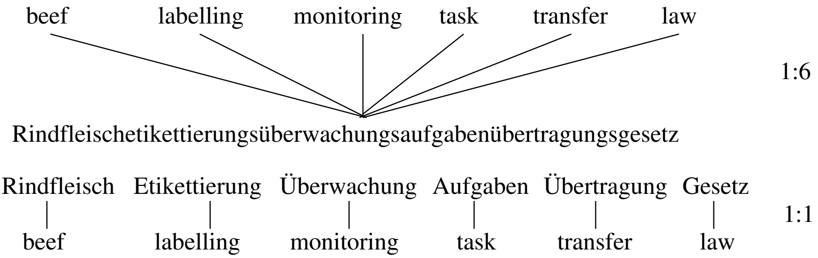
Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

1:6

Rindfleisch Etikettierung Überwachung Aufgaben Übertragung Gesetz

This is a real example!

Compound Splitting for SMT



beef labelling monitoring task transfer law

1:6

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

Rindfleisch Etikettierung Überwachung Aufgaben Übertragung Gesetz

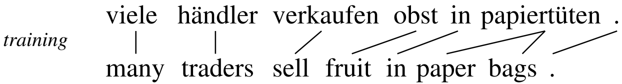
1:1

beef labelling monitoring task transfer law

This is a real example!

→ use a morphological analyser to spllit compounds
→ disambiguate analyses using corpus frequencies
→ no modification of the SMT system
→ instead modify the training data!

German to English SMT Example

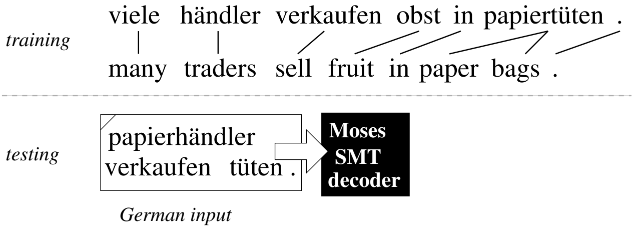


training

viele händler verkaufen obst in papiertüten .

many traders sell fruit in paper bags .

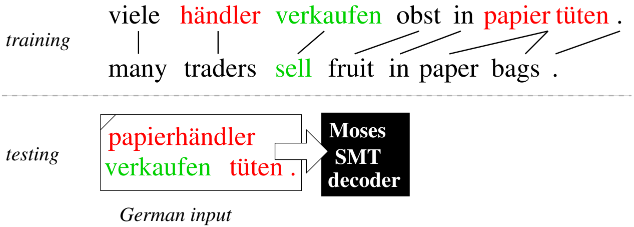
German to English SMT Example



training
viele händler verkaufen obst in papiertüten .
many traders sell fruit in paper bags .
testing
papierhändler verkaufen tüten .
Moses SMT decoder
German input

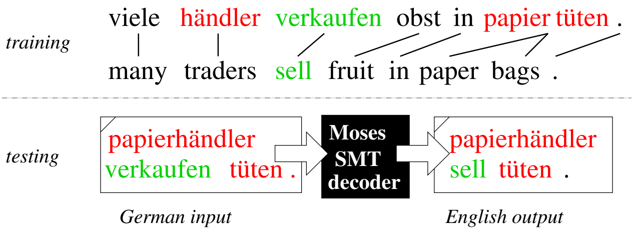
Baseline

German to English SMT Example


training
viele händler verkaufen obst in papier tüten .
many traders sell fruit in paper bags .
testing
papierhändler
verkaufen tüten .
Moses
SMT
decoder
German input

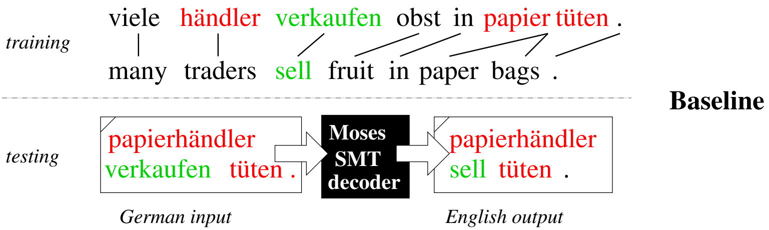
Baseline

German to English SMT Example


training
viele händler verkaufen obst in papier tüten .
many traders sell fruit in paper bags .
testing
papierhändler verkaufen tüten .
Moses SMT decoder
papierhändler sell tüten .
German input
English output

Baseline

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

Baseline

testing

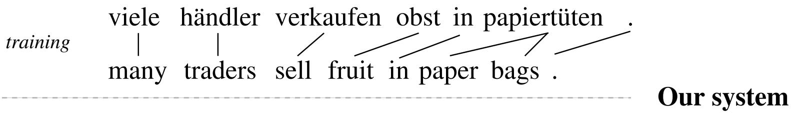
papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output



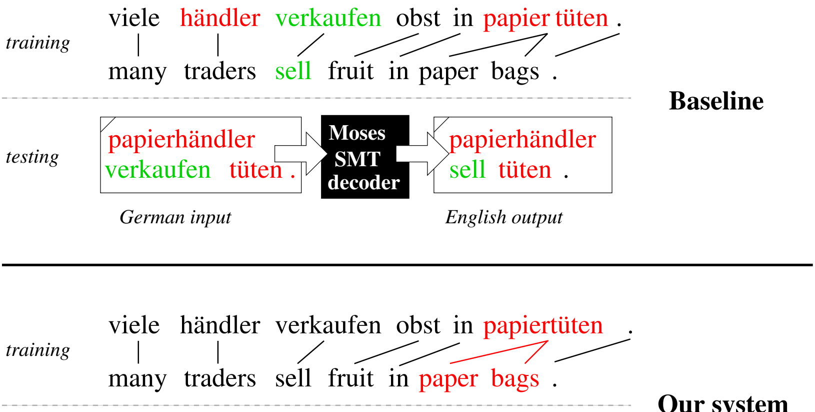
training

viele händler verkaufen obst in papiertüten .

many traders sell fruit in paper bags .

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

Baseline

testing

papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output

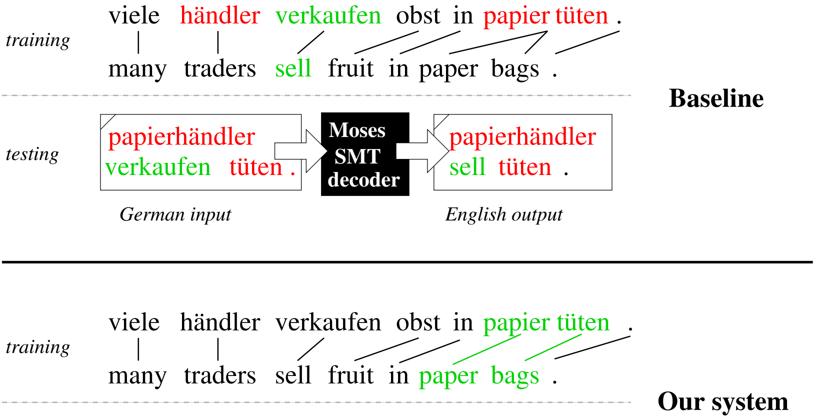
training

viele händler verkaufen obst in papiertüten .

many traders sell fruit in paper bags .

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

Baseline

testing

papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output

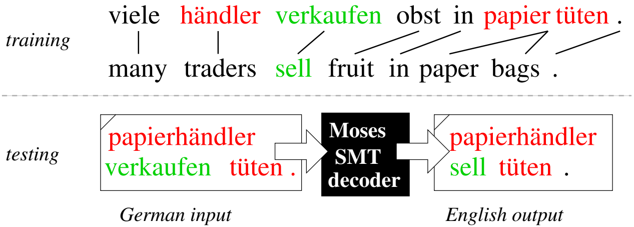
training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

papierhändler verkaufen tüten .

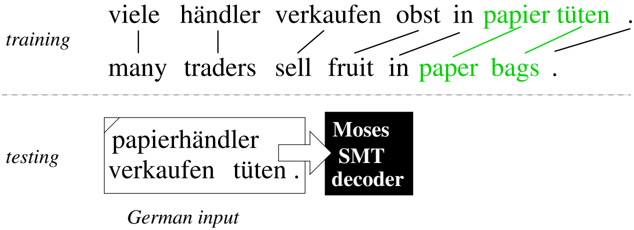
Moses SMT decoder

papierhändler sell tüten .

German input

English output

Baseline



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

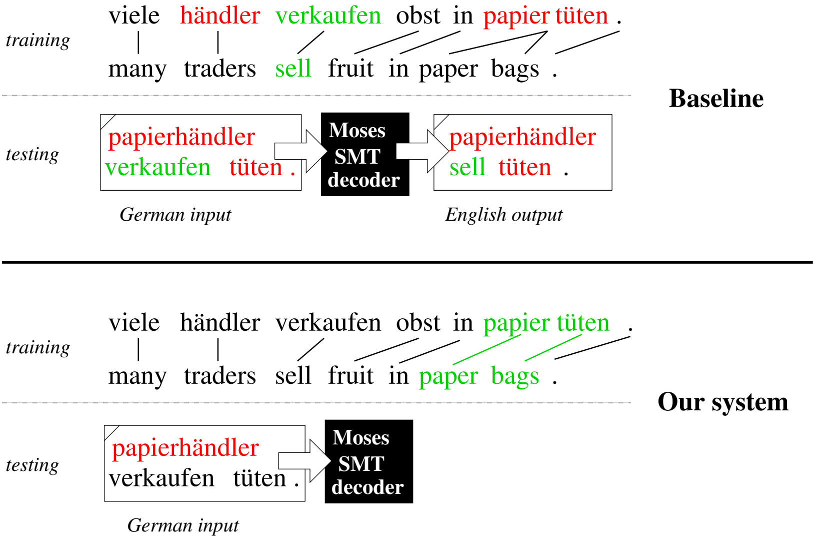
papierhändler verkaufen tüten .

Moses SMT decoder

German input

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output

Baseline

training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

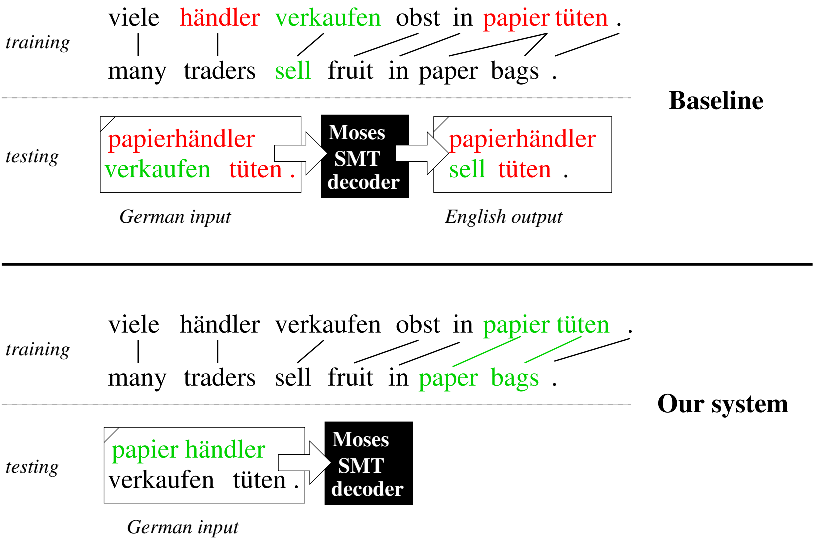
papierhändler verkaufen tüten .

Moses SMT decoder

German input

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output

Baseline

training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

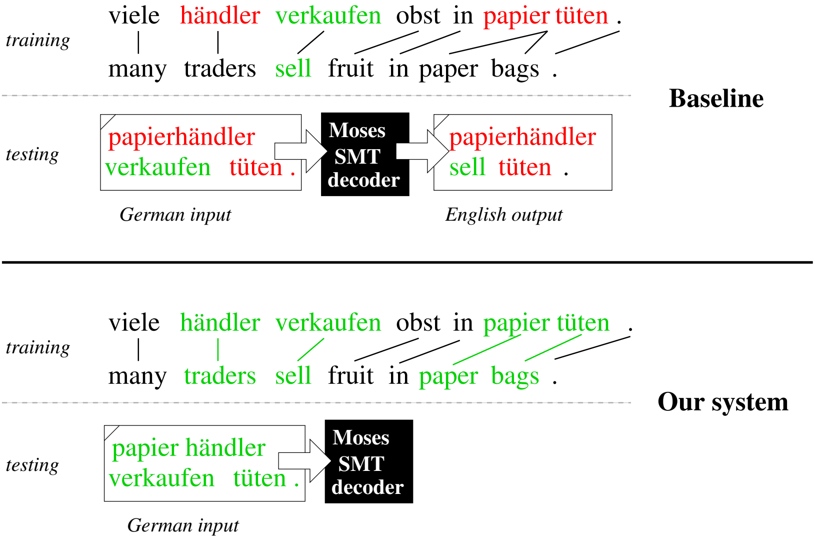
papier händler verkaufen tüten .

Moses SMT decoder

German input

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output

Baseline

training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

testing

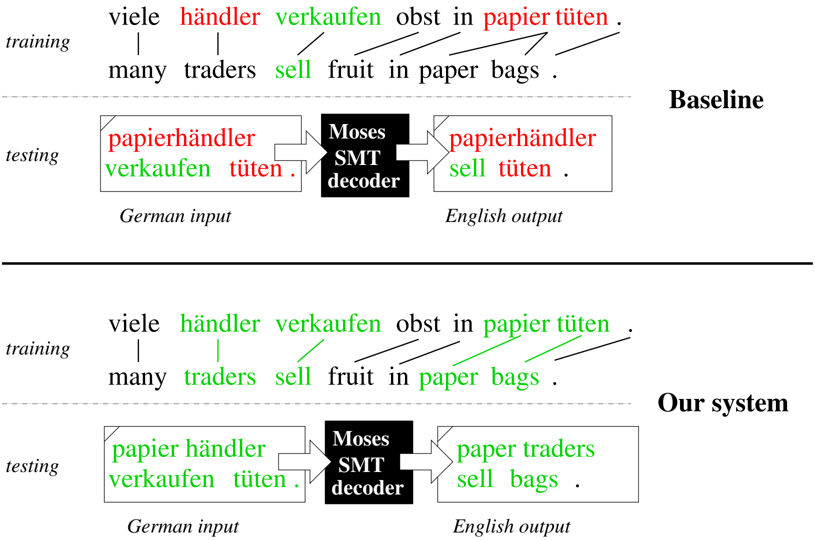
papier händler verkaufen tüten .

Moses SMT decoder

German input

Our system

German to English SMT Example



training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

Baseline

testing

papierhändler verkaufen tüten .

Moses SMT decoder

papierhändler sell tüten .

German input

English output

training

viele händler verkaufen obst in papier tüten .

many traders sell fruit in paper bags .

Our system

testing

papier händler verkaufen tüten .

Moses SMT decoder

paper traders sell bags .

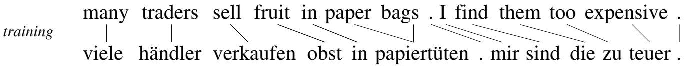
German input

English output

Now: opposite translation direction!!!

Pay Attention You Must!!

English to German SMT Example

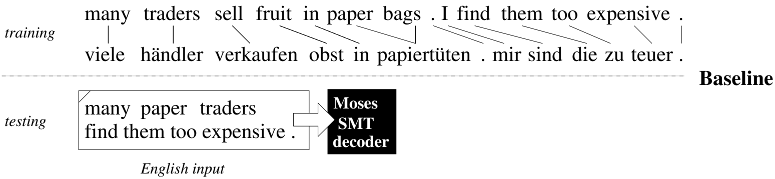


training

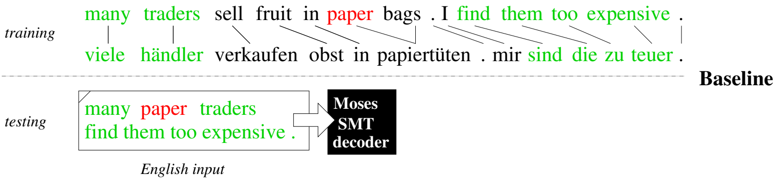
many traders sell fruit in paper bags . I find them too expensive .

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

English to German SMT Example


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

English to German SMT Example



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

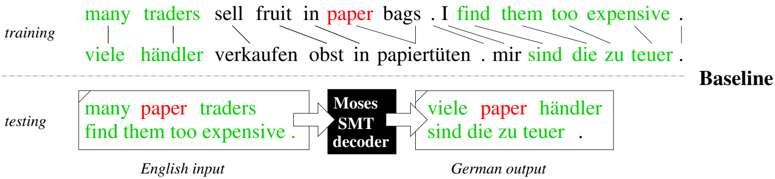
testing

many paper traders find them too expensive .

Moses SMT decoder

English input

English to German SMT Example



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 .

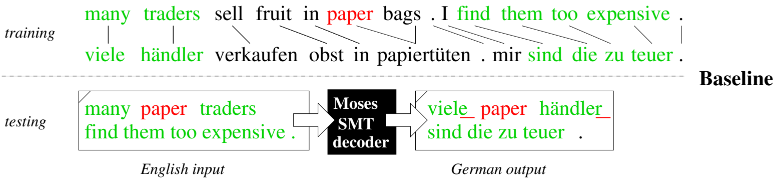
Moses SMT decoder

viele paper händler sind die zu teuer .

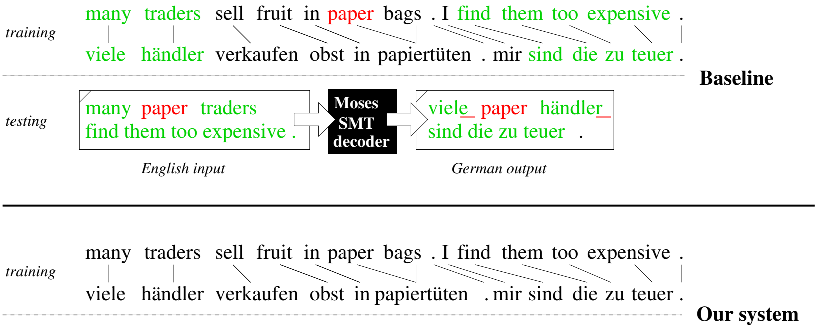
English input

German output

English to German SMT Example


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

English to German SMT Example



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

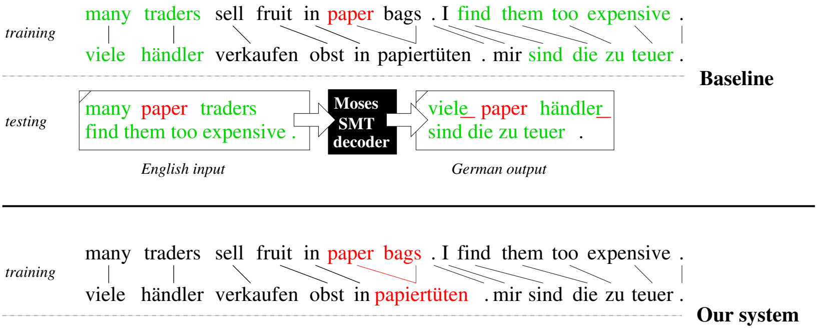
training

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

English to German SMT Example



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

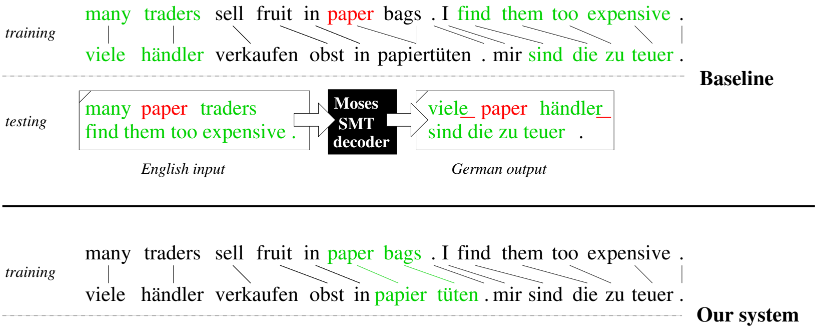
training

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

English to German SMT Example



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

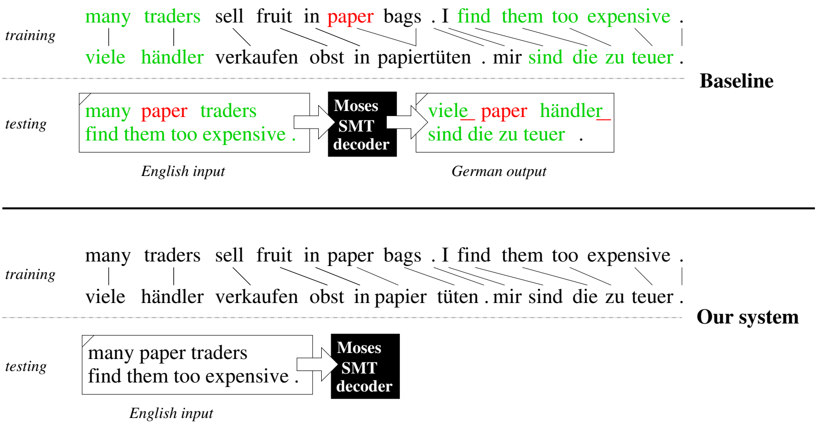
training

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

viele händler verkaufen obst in papier tüten . mir sind die zu teuer .

Our system

English to German SMT Example



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 papier tüten . mir sind die zu teuer .

Our system

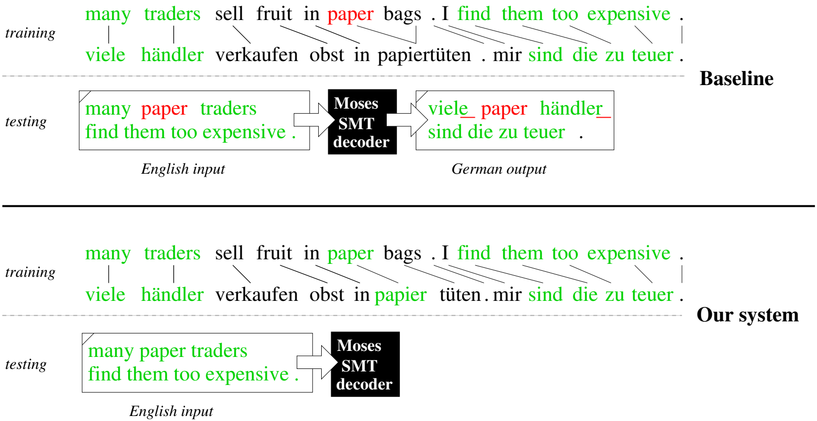
testing

many paper traders
find them too expensive .

English input

Moses SMT decoder

English to German SMT Example



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 papier tüten . mir sind die zu teuer .

Our system

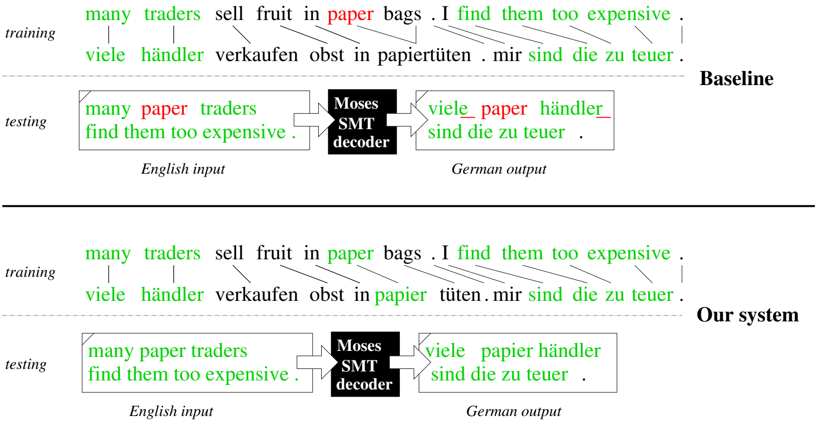
testing

many paper traders
find them too expensive .

English input

Moses SMT decoder

English to German SMT Example



training

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

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

testing

many paper traders find them too expensive .

Moses SMT decoder

viele_ paper händler_ sind die zu teuer .

English input

German output

Baseline

training

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

viele händler verkaufen obst in papier tüten . mir sind die zu teuer .

testing

many paper traders find them too expensive .

Moses SMT decoder

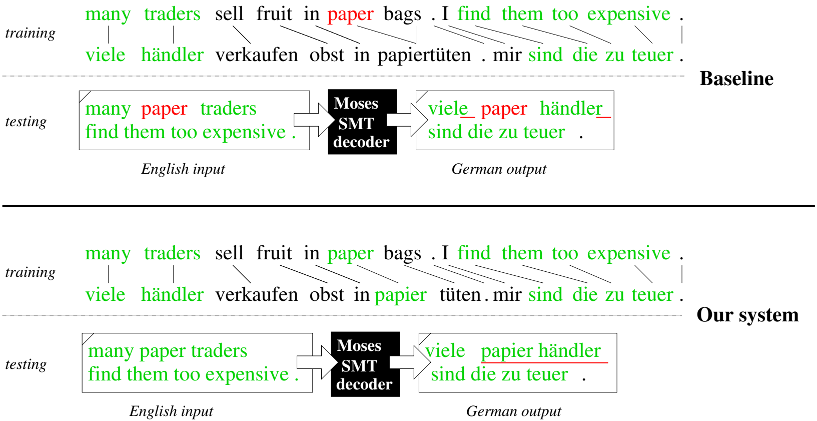
viele papier händler sind die zu teuer .

English input

German output

Our system

English to German SMT Example



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 papier tüten . mir sind die zu teuer .

Our system

testing

many paper traders find them too expensive .

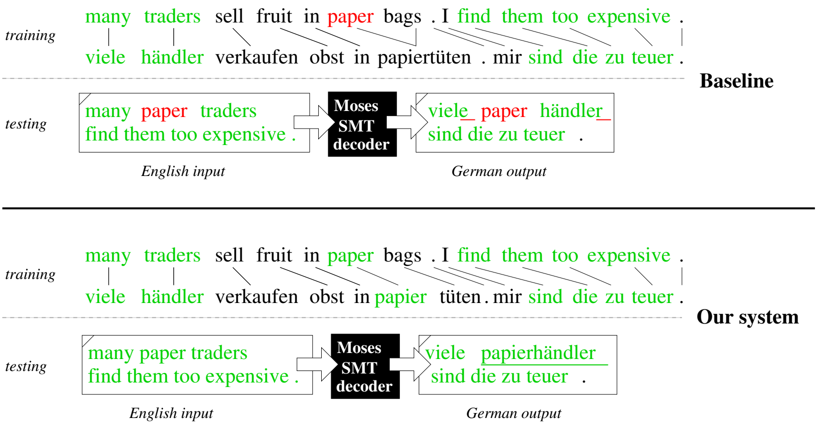
Moses SMT decoder

viele papier händler sind die zu teuer .

English input

German output

English to German SMT Example



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 papier tüten . mir sind die zu teuer .

Our system

testing

many paper traders find them too expensive .

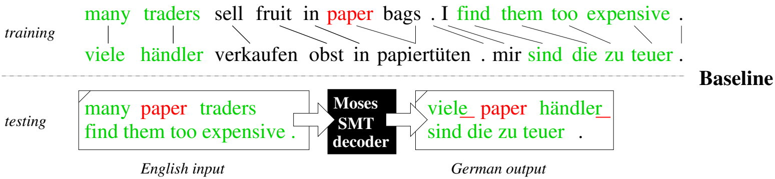
Moses SMT decoder

viele papierhändler sind die zu teuer .

English input

German output

English to German SMT Example



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

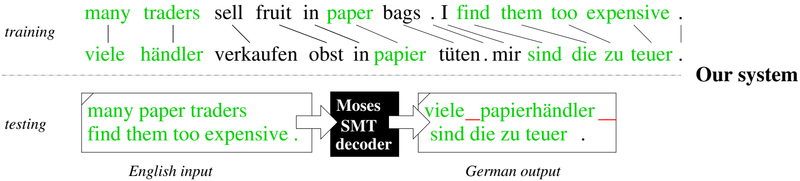
many paper traders find them too expensive .

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 papier tüten . mir sind die zu teuer .

Our system

testing

many paper traders find them too expensive .

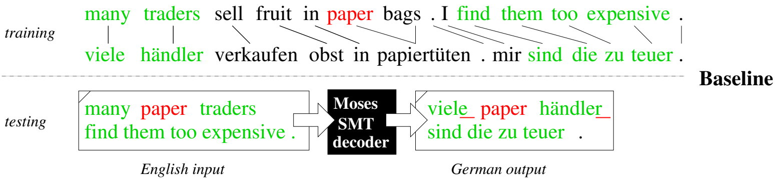
Moses SMT decoder

viele_papierhändler _ sind die zu teuer .

English input

German output

English to German SMT Example



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

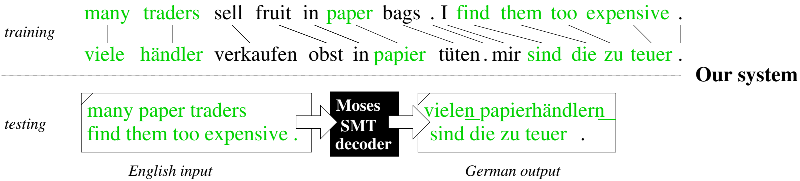
many paper traders find them too expensive .

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 papier tüten . mir sind die zu teuer .

Our system

testing

many paper traders find them too expensive .

Moses SMT decoder

vielen_papierhändlern_ sind die zu teuer .

English input

German output

German to English SMT Example

Compound Processing....

- allows to translate compounds that have not occurred in the training data:
  - provided that they have been properly split
  - their parts must have occurred in the training data
  - it is irrelevant how the parts occurred:
    as simplex words, compound modifiers or heads
- enhances the word counts of simplex words and thus makes their translations more reliable as well
- can produce unseen inflectional variants of seen words
- can produce coherent inflected sequences of words

How to use identified MWEs to improve translation

	MWE characteristics	Solution?
1	multiple lexical units	Compound processing
2	many types, not many tokens	Lemmatisation
3	semantically non compositional	**Tell SMT where MWEs are**
4	discontinuous components	**Tell SMT where MWEs are**

Support Verb Constructions in SMT

- Support-verb constructions (SVCs):
  semantically light verb with a predicative noun
- The verb neither contributes its full meaning,
  nor is it completely void
  - ***to take*** *a bath*
  - ***to take*** does not contribute its full meaning,
    but is different from ***to make a bath***
- SVCs often close in meaning to a corresponing full verb

	English	
V+NP	make a contribution	contribute
V+PP	take into account	consider
	German	
V+NP	einen Beitrag leisten lit. *a contribution achieve*	beitragen *to contribute*
V+PP	in Frage stellen lit. *in question put*	hinterfragen *to question*

⇐ we focus on V+NPs

Support Verb Constructions in SMT

- Translations are learned from **word-aligned parallel data**, text is considered a **sequence of words**
  → MWEs are not distinguished from any other sequences
- Target-side **language model** provides context information
- However: SMT systems often choose the **default translation** regardless of the context

  vertreten → *to represent*

  in the context of ***the Auffassung vertreten***

vertreten$_{SVC}$ → *to represent*	⇒	**to represent the view*
vertreten$_{SVC}$ → *to take*	⇒	*to take the view*

⇒ Default translation is often wrong in the case of an SVC

⇒ Marking the **verb of the SVC in the training data** so that the system learns the different translations

Non-adjacent Support Verb Constructions

- An SVC where **noun and verb are adjacent** is likely to be correctly **translated as one phrase**
- Much more difficult for an **isolated verb**: no connection to the noun → likely to be translated with the default translation

- Relatively free word order in German
- Often large **gaps between verbs and nouns** in German

Dazu **leistet**$_V$ die Effizienz des Vermittlungsverfahrens einen substanziellen **Beitrag**$_N$.
To that **make** *the effectiveness of the codecision procedure a substantial* **contribution**.

The effectiveness of the codecision procedure has **made** a substantial **contribution** in this case .

⇒ Identification of SVCs requires parsed data

Related Work

- **Static approach:**
  modification of training and test data
- **Dynamic approach:**
  add MWE-based features to phrase-table

Related Work

- Carpuat and Diab (2010)
  - static: merge MWEs into one phrase
  - dynamic: add count-based features for MWEs into phrase-table
  - using lexical MWE resources (English)
  - comparable results for static and dynamic approaches
- Cholakov and Kordoni (2014)
  - Model phrasal verbs in English-Bulgarian SMT
  - dynamic: adding linguistic features to phrase-table
  - better results for dynamic approach
- Relation to previous work
  - we apply the static approach
  - no merging of the parts of an MWE to form a single unit: we only mark the verb of an SVC

Procedure

SMT from German into English

(1) Extraction of verb-object pairs (lemma-level) from dependency-parsed data

(2) Identification of SVCs using associaton measures

(3) Creation of several SVC sets with different degrees of idiomaticity

(4) Markup for verbs in SVCs in the training data for the training data for the SMT system

(5) Training of the SMT system using standard settings and translation of test-set

SVC-sets with different degrees of idiomaticity

Investigate **different thresholds** of log-likelihood scores for the ranked list of verb-object pairs

⇒ obtain different sets with **varying degrees of idiomaticity**

	training		testing	
	types	token	types	token
all	30,6572	1,102,166	794	881
freq≥5[1]	25,610	713,734	461	537
LL ≥ 1000	338	181,818	58	94
LL ≥ 500	693	240,369	95	139
LL ≥ 350	1,024	271,908	120	168
LL ≥ 250	1,473	304,148	142	191

Table: Number of SVCs in the training data and test set

[1] Verb-object pairs with a frequency≤5 are excluded

Verb Markup

- For each of the verb-object pairs in the subsets: **mark the verbs occuring within an SVC** in the training and test data
- Independent verbs with a literal sense are distinct from verbs with an idiomatic meaning

⇒ helps the SMT system to distinguish these verbs

SVC	Das hat einen wichtigen **Beitrag geleistet_SVC**.
	This has an important ***contribution made****.*
	This has **made** an important **contribution**.
other	Ich glaube , dass sie sehr viel Gutes **geleistet** hat .
	I believe, that it very much good ***achieved*** *has.*
	I believe that it has **achieved** a great deal of good .

- Lemmatized list of SVCs, inflected forms in training/test data

Results

BLEU: measures n-gram similarity to one human reference translation

Experiments	BLEU
Baseline	20.49
Exp1000	21.01
Exp500	21.01
Exp350	20.89
Exp250	20.84

Table: BLEU scores on the WMT 2014 testset.

Improved Verb Translations

- In addition to BLEU: investigate the **translation of verbs**
- **Missing verbs** are a typical problem in DE–EN translation
- Verbs play a primary role in understanding a sentence → missing verbs have a **severe effect on translation quality**

System	# sentences with at least one full verb
Baseline	2,378
Exp1000	2,412
Exp500	2,413
Exp350	2,412
Exp250	2,411
Reference	2,712

⇒ Each system produces more verbs compared to the baseline

Sentence-level Verb Comparison

- Comparison of **verb translations** with the **reference translations** (lemma-level matching)

Lemma-level verb match count	
Baseline matches reference	3,505
Exp250 matches reference	**3,648**

⇒ System yields more verbs that match the reference translation

Success Stories (1)

Baseline: no verb translation
Exp250: correct translation of the SVC verb

input	Sie wollen herausfinden, welche **Rolle** der Riesenplanet bei der Entwicklung des Sonnensystems **gespielt** hat. *They wanted to find out, what* **role** *the giant-planet for the development of-the solar-system* **played** *has.*
reference	They want to find out what **role** the giant planet has **played** in the development of the solar system.
baseline	You want to find out what **role** the *Riesenplanet* in the development of the solar system.
Exp250	They want to find out what **role** the *Riesenplanet* **played** in the development of the solar system.

Success Stories (2)

Baseline: default translation of the verb
Exp250: SVC translation of the verb

input	“Ich **vertrete** die **Auffassung**, dass eine hinreichende Grundlage für eine formelle Ermittlung besteht“, sagte er. *I **take** the **view** that a sufficient basis for a formal investigation exists, said he.*
reference	“I **am** of the **opinion** that a sufficient basis exits” for a formal investigation, he said.
baseline	„I **represent** the **view** that a sufficient basis for a formal investigation is", he said.
Exp250	„I **take** the **view** that a sufficient basis for a formal investigation is", he said.

Effect on Translation Probabilities

Comparison of translation options for different uses of *treffen*

Baseline		**Exp1000**			
treffen		*treffen*		*treffen_SVC*	
prob	transl.	prob	transl.	prob	transl.
0.295	meeting	0.315	meeting	0.237	take
0.105	meetings	0.112	meetings	0.176	make
0.086	take	0.074	take	0.032	will
0.059	make	0.048	make	0.022	decide
0.036	meet	0.039	meet	0.019	taken
0.013	be	0.012	be	0.012	reach
0.011	hit	0.012	hit	0.009	will take
0.010	affect	0.011	adopt	0.009	will make
0.010	adopt	0.011	affect	0.009	to take
0.007	taken	0.007	taken	0.009	to make

Ongoing Work: Verb Markup with Nouns

- Different SVCs share the same verb

	literal	idiomatic
Maßnahmen **ergreifen**	"to grasp measures"	*to take measures*
Flucht **ergreifen**	“to grasp escape”	*to escape*
Wort **ergreifen**	“to grasp (the) word”	*to rise to speak*

- Explicitely distinguish verb translations of different SVCs

Er kann ein paar technische **Maßnahmen ergriffen_**SVC_Maßnahme werden.
he can **take** additional technical **measures**.

Delegierte dürfen nicht mehr als ein Mal ... das **Wort ergreifen_**SVC_Wort.
No delegate shall be allowed to **speak** more than once ...

Merci

IMS, University of Stuttgart

Marion Di Marco (née Weller)
Alex Fraser (now: CIS, Munich)
Sabine Schulte im Walde
Manju Nirmal
Ulrich Heid (now: Uni Hildesheim)

To Thank You I Want

To thank you I want

References

- **Marine Carpuat and Mona Diab (2010)**, *Task-based evaluation of multiword-expressions: a pilot study in statistical machine translation* in NAACL'10: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.
- **Konstadin Cholakov and Valia Kordoni (2014)**, *Better statistical machine translation through linguistic treatment of phrasal verbs* in EMNLP'14: Proceedings of the Conference on Empirical Methods in Natural Language Processing.
- **Helena de Medeiros Casel, Carlos Ramisch, Maria das Gracas Volpe Nunes and Aline Villavicenci (2009)**, *Alignment-based extraction of multiword expressions* in Language Resources and Evaluation Special Issue on Multiword expression: hard going or plain sailing, Springer.
- **Fabienne Fritzinger (2008)**, *Extracting Multiword Expressions from Parallel Text*, Master's thesis, University of Stuttgart

. • **Yulia Tsvetkov and Shuly Wintner (2011)**, *Identification of multi-word expressions by combining multiple linguistic information sources* in EMNLP'11: Proceedings of the Conference on Empirical Methods in Natural Language Processing.

- **Begona Villada Moiron and Jörg Tiedemann (2006)**, *Identifying idiomatic expressions using automatic word-alignment* in EACL'2006: Proceedings of the EACL 2006 Workshop on Multiword-expressions in a multilingual context.