

Extending Memory-Based Machine Translation to Phrases

Maarten van Gompel, Antal van den Bosch, Peter Berck

February 2010



What is Phrase-based Memory-based Machine Translation?

1 Machine translation

- Data-driven
- Form of Example Based MT

2 Memory-based

- Lazy-learning classifier maps fragments in the source language to fragments in the target language

3 Phrase-based

- New!
- Taking phrases of variable length

Previous research in MBMT: Training

① Using an aligned parallel corpus

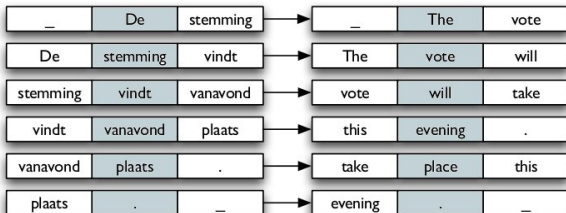
De stemming vindt vanavond plaats .
The vote will take place this evening .

Previous research in MBMT: Training

1 Using an aligned parallel corpus

De stemming vindt vanavond plaats .
 The vote will take place this evening .

2 Generate fragments, trigram-to-trigram mapping

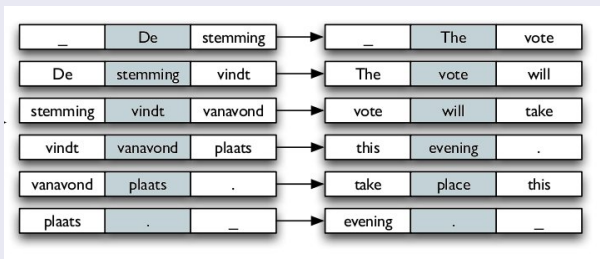


Previous research in MBMT: Training

- Using an aligned parallel corpus

De stemming vindt vanavond plaats .
 The vote will take place this evening .

- Generate fragments, trigram-to-trigram mapping



- Construct a losslessly compressed decision tree (training)

Previous research in MBMT: Testing & Decoding

- 1 Extract trigram fragments from test data.



Previous research in MBMT: Testing & Decoding

- 1 Extract trigram fragments from test data.



- 2 Pass fragments through classifiers to obtain translations



Previous research in MBMT: Testing & Decoding

- 1 Extract trigram fragments from test data.



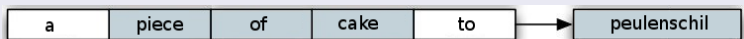
- 2 Pass fragments through classifiers to obtain translations



- 3 **Decoder:** Re-assemble translated fragments, making use of overlap in context

Extending MBMT to phrases

Variable length phrases instead of trigrams:



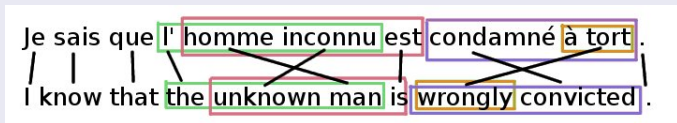
Map phrases *in context* from source language to target language.

Key questions

- 1 Does a phrase-based approach provide better results?
- 2 *How to extract phrases?*
- 3 How to decode and reassemble fragments?

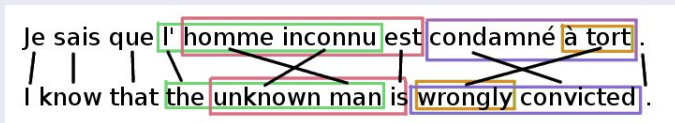
Phrase Extraction

How to extract phrases?



Phrase Extraction

How to extract phrases?

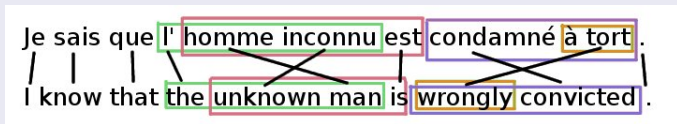


Three methods:

- 1 Phrase-translation table (generated by Moses)

Phrase Extraction

How to extract phrases?

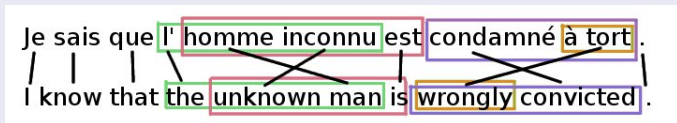


Three methods:

- 1 Phrase-translation table (generated by Moses)

Phrase Extraction

How to extract phrases?

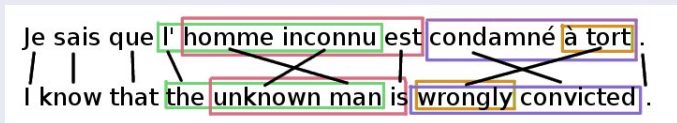


Three methods:

- 1 Phrase-translation table (generated by Moses)
- 2 Phraselist (common n-grams)

Phrase Extraction

How to extract phrases?



Three methods:

- 1 Phrase-translation table (generated by Moses)
- 2 Phraselist (common n-grams)
- 3 Marker based Chunking

The rapporteur	has	also quite rightly stated	that	parliament was not heard	in	time	regarding	the guidelines		
↑	↑		↑		↑		↑			
De rapporteur	heeft	ook zeer terecht gezegd	dat	het parlement	niet	tijdig	over	de voorschriften	is	gehoord
↑	↑		↑		↑		↑		↑	

Decoder - Fragmentations

A test sentence may be fragmented in various ways:

Het

boek

ligt op

de tafel

Het boek

ligt op

de

tafel

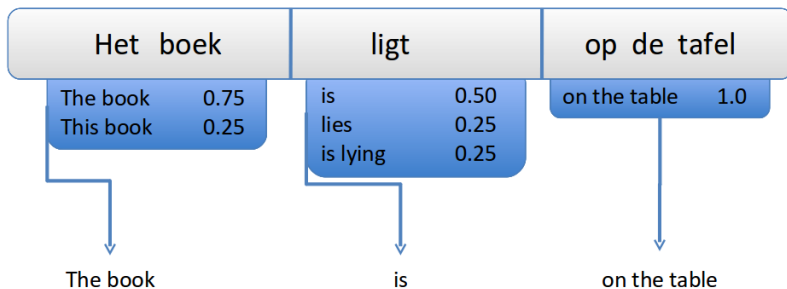
Het boek

ligt

op de tafel

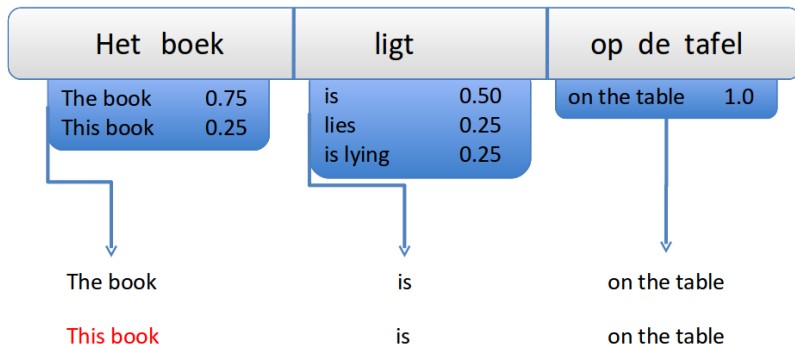
Decoder - Fragmentation

For each fragment in a fragmentation, the classifier predicts translations, an initial hypothesis is constructed:



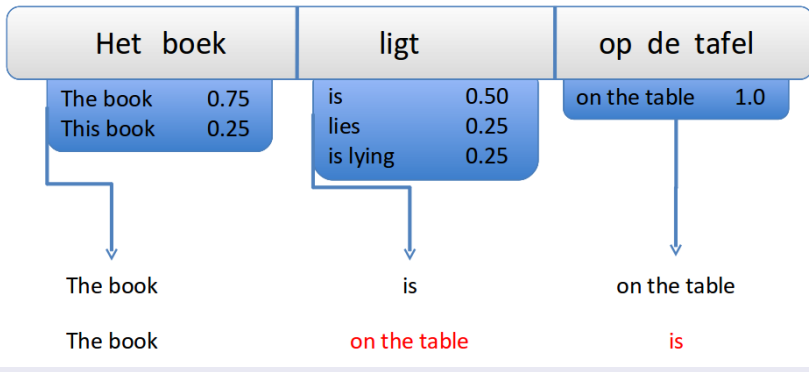
Decoder - Substitution

Substitution operation - A translation is substituted for another



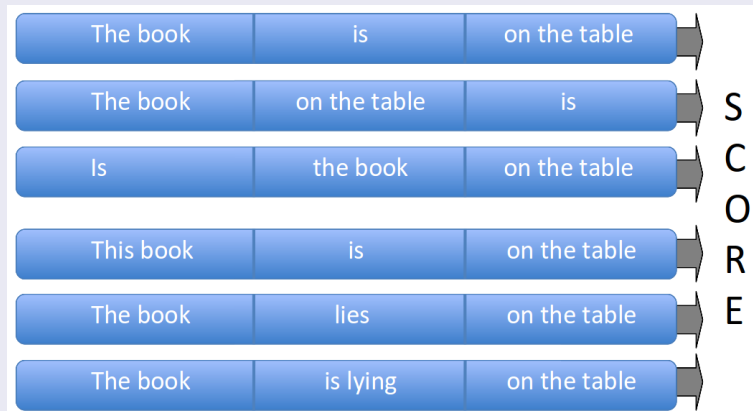
Decoder - Swap

Swap operation - Two fragments swap places



Decoder

- Local beam search
- Hypotheses are scored, and hypotheses from multiple fragmentations recombined



Data and Setup

- **Language pair:** Dutch to English
- **Parallel corpora:** OpenSubtitles & EMEA
- **Training data:** 286,160 resp. 187,189 sentences.
- **Test data:** 1000 sentences.
- Simple parameter optimisation on decoder (on separate trial sets)
- Evaluation using BLEU, NIST, METEOR, WER & PER metrics.

Results compared to MBMT (OpenSubtitles & EMEA)

Decoder	Extract	Classif.	BLEU	NIST	METEOR	WER	PER
MBMT			0.1631	4.243	0.3835	68.39	61.33
CSIMT			0.2002	4.750	0.4431	68.42	55.18
Wb-PB			0.2163	5.136	0.4644	55.23	48.22
PB	Phr.tbl	single	0.2300	5.055	0.4623	54.47	49.18

Decoder	Extract	Classif.	BLEU	NIST	METEOR	WER	PER
MBMT			0.2533	5.115	0.4801	72.78	63.66
CSIMT			0.3013	5.938	0.5333	63.00	50.85
Wb-PB			0.2715	5.600	0.5381	65.99	57.25
PB	Phr.tbl	multi	0.3078	6.019	0.5449	58.76	51.63

Results compared to state-of-the art systems

Decoder	Extract	Classif.	BLEU	NIST	METEOR	WER	PER
Moses			0.3289	5.9035	0.5408	53.29	46.96
Google			0.3056	5.7893	0.5224	50.10	45.08
PB	Phr.tbl	single	0.2300	5.0550	0.4623	54.47	49.18
Systran			0.1749	4.5828	0.4500	60.77	54.61

Decoder	Extract	Classif.	BLEU	NIST	METEOR	WER	PER
Moses			0.4701	7.0593	0.6501	46.55	39.36
Google			0.3918	6.3772	0.5830	57.57	50.44
PB	Phr.tbl	multi	0.3078	6.019	0.5449	58.76	51.63
SysTran			0.2895	5.4716	0.5366	63.24	55.14

Research Question

Does a phrase-based approach improve MBMT?

Research Question

Does a phrase-based approach improve MBMT?

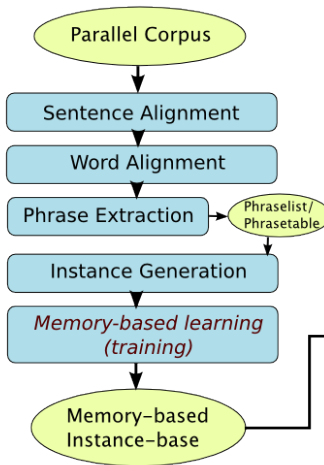
Conclusions

- Improvement over previous MBMT methods, but only with phrase-table method
- Phrase-based improvement over word-based baseline smaller than expected
- Predicting classes without context information yields better results

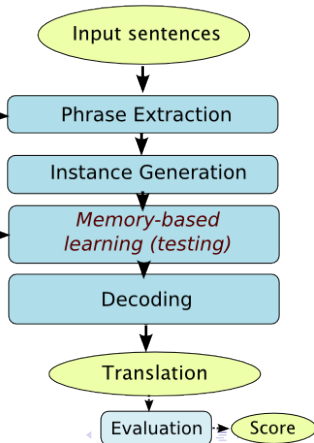


Questions?

TRAINING:



TESTING:



Setup

No	Corpus	System	BLEU	NIST	METEOR	WER	PER
1	OpenSub	Moses	0.3289	5.903	0.5408	53.29	46.96
2	OpenSub	Google	0.3056	5.790	0.5224	50.1	45.08
3	OpenSub	PBMBMT	0.2314	5.061	0.4630	54.44	49.17
4	OpenSub	Systran	0.1749	4.583	0.4500	60.77	54.61
1	EMEA	Moses	0.4701	7.059	0.6501	46.55	39.36
2	EMEA	Google	0.3918	6.377	0.5830	57.57	50.44
3	EMEA	PBMBMT	0.3084	6.015	0.5460	59.03	52.02
4	EMEA	Systran	0.2895	5.472	0.5366	63.24	55.14

Table: A comparison with state-of-the-art systems

Decoder	Extract. Meth.	Instance F.	BLEU	NIST	METEOR	WER	PER
MBMT	-	-	0.148	3.817	0.3835	67.76	62.03
CSIMT	-	-	0.2002	4.750	0.4431	68.42	55.18
PBMBMT	-	-	0.2163	5.136	0.4644	55.23	48.22
PBMBMT	phrase table	split-files	0.2256	5.004	0.4583	55.28	49.74
PBMBMT	phrase table	fixed-feat.	0.2300	5.055	0.4623	54.47	49.18
PBMBMT	phrase table	single-feat.	0.1142	3.026	0.3201	72.81	68.28
PBMBMT	phrase list (5)	split-files	0.2152	4.798	0.4445	54.57	49.89
PBMBMT	phrase list (25)	split-files	0.2184	4.975	0.4529	54.09	48.79
PBMBMT	phrase list (25)	fixed-feat.	0.2190	4.980	0.4543	54.09	48.77
PBMBMT	marker-based	split-files	0.1394	3.360	0.3437	66.40	62.38
PBMBMT	marker-based	fixed-feat.	0.1003	2.935	0.3057	76.79	71.16

Table: Main results on the **OpenSubtitles** corpus, Dutch to English

Decoder	Extract. Meth.	Instance F.	BLEU	NIST	METEOR	WER	PER
CSIMT	-	-	0.3013	5.938	0.5333	63.00	50.85
PBMBMT	-	-	0.2715	5.600	0.5381	65.99	57.25
PBMBMT	phrase table	split-files	0.3078	6.019	0.5449	58.76	51.63
PBMBMT	phrase table	fixed-feat.	0.3075	6.011	0.5455	59.00	52.02
PBMBMT	phrase list (25)	split-files	0.2440	5.378	0.4967	62.82	56.86
PBMBMT	phrase list (25)	fixed-feat	0.2440	5.352	0.4946	62.74	56.67
PBMBMT	marker-based	split-files	0.2370	4.612	0.4513	74.37	66.78

Table: Main results on the **EMEA** corpus, Dutch to English

	BLEU	NIST	METEOR	WER	PER
Without context (0X0)	0.2184	4.9748	0.4529026	54.0885	48.7859
With context 1 (1X1)	0.1211	3.3015	0.3493939	65.6648	61.5435

Table: A comparison of usage of context in classes, tested on the OpenSubtitles corpus (phrase list 25 extr., split-files format)

n	phrase table	phrase list (25)	marker-based chunking
1	1847887	1847887	2047875
2	1213597	619311	131886
3	845796	109338	118021
4	595531	13200	64915
5	420294	1357	31777
6	283107	326	13966
7	-	-	5779
8	-	-	2036
9	-	-	720
Total	5206212	2591419	2416975

Table: Overview of the number of phrases extracted for different phrase-lengths (determined using the split-file approach on the OpenSubtitles corpus)