Integration of a Multilingual Preordering Component into a Commercial SMT Platform

Anita Ramm, Riccardo Superbo, Dimitar Shterionov,
Tony O'Dowd, Alexander Fraser

IMS, University of Stuttgart
KantanMT.com
CIS, University of Munich

May 29th, 2017

Outline

1. Word Order in SMT
2. Our Approach
3. Experiments and Evaluation
   - SMT
   - NMT
4. Summary and Future Work

Outline

Word Order in SMT

Introduction

Word order in Statistical Machine Translation

- Translation of a SL word needs to be placed in a TL-specific position

⇒ Fluency of the translations

- Problematic for languages with considerably different syntactic structure, e.g.:

EN:	...that	I	read	a	book	yesterday		.
DE:	...dass	ich	gestern	ein	Buch	gelesen$_{part}$	habe$_{finV}$	.

- Particularly, the long-range reorderings lead to **false placement or omission** of TL words
  ⇒ Negative impact on the adequacy and fluency of the generated translations!

Reordering in SMT

- Many different approaches to handle reordering problems within SMT (cf. **[Bisazza & Federico, 2016]**)
- One of the simplest, yet most effective approaches: **preordering**

EN: ...that I a book yesterday read .

DE: ...dass ich gestern ein Buch gelesen$_{part}$ habe$_{finV}$.

- Reordering may be based on either **automatically derived rules** or **hand-crafted rules**
- Reordering is carried out on **parsed SL sentences**, on the **POS-tagged** or **non-processed SL corpus**

Outline

Reordering for a commercial SMT platform

- Multilingual preordering component for English→Japanese/German/Chinese SMT
- Rule-based reordering based on hand-crafted rules
  ⇒ **improvement already shown by previous research**
  ⇒ **easily adapt to the client's needs**
- Stand-alone component used as a part of corpus preprocessing
  ⇒ **ensures backward compatibility**
- Generic implementation independent of the used parsing software
  ⇒ **allows for usage of different parsers**

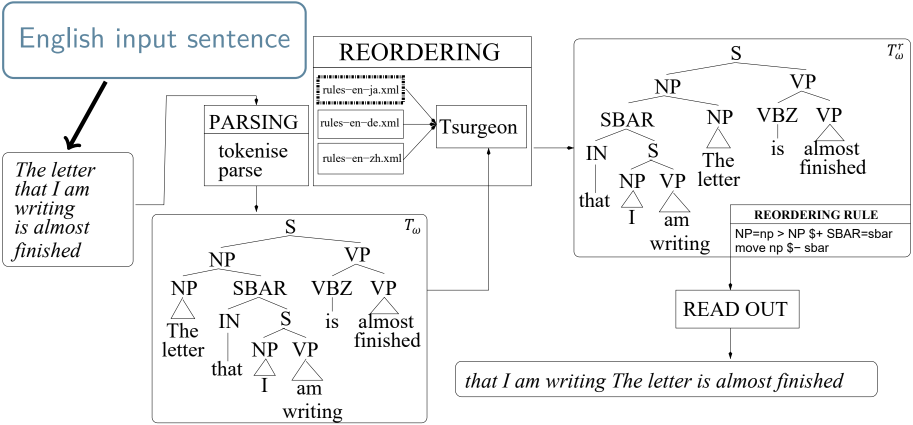
Our Approach

Languages under consideration

1. English→German:
   - Rule set based on [Gojun & Fraser, 2012]
   - Movements of the verbs and negation to capture different positions of the verbs between EN and DE
   - Total of 9 reordering rules
2. English→Japanese:
   - Rule set based on [Lee et al., 2012]
   - Movements of various sentence constituents
   - Total of 7 reordering rules
     and 1 insertion rule (to cope with null subjects in JA)
3. English→Chinese:
   - Rule set developed based on [Wang et al., 2007] and [Wu, 2016]
   - Movements of various sentence constituents
     ⇒ doesn't work!
   - Exclusive reordering of NP-PPs and ofPPs
     (2 rules apply independently)

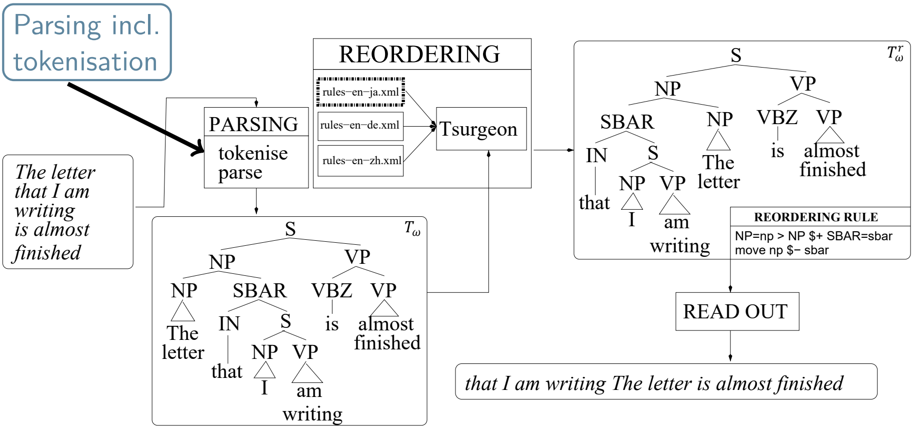
Our Approach

Architecture overview



Our Approach

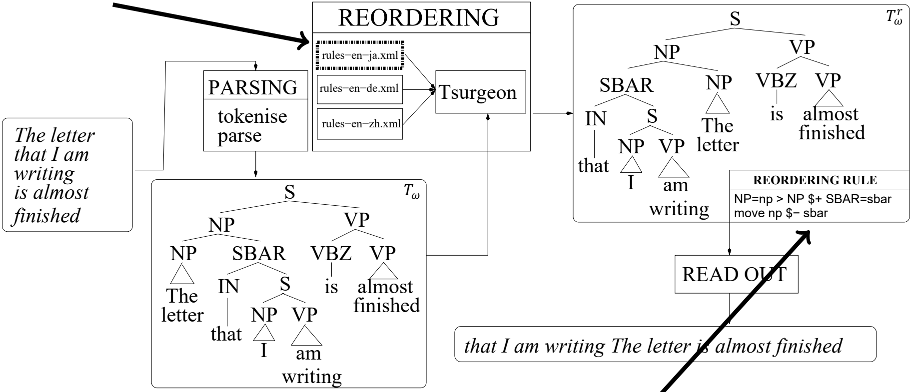
Architecture overview



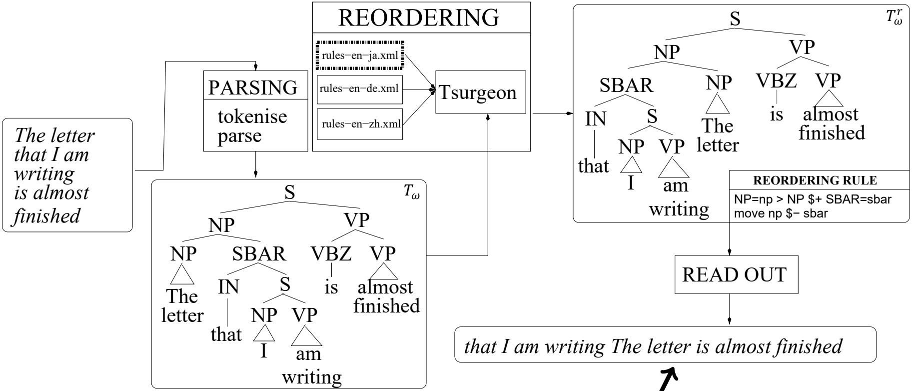
Our Approach

Architecture overview

Reordering based on Tsurgeon



and hand-crafted tree modification rules



Read out the reordered sentence

Our Approach

Architecture overview

Processing steps

- For each sentence in the input file:
  1. Tokenise/parse the input sentence
     (constraints: 5<len(sentence)<60, no special characters)
  2. Modify the parse tree (i.e., perform reordering)
  3. Read out the sentence from the modified parse tree
- After processing the entire input file,
  continue with tokenisation, lowercasing, etc.

Reordering

- **Tsurgeon-based modifications** of the parse trees
  [Levy & Andrew, 2006]
- Rules are defined following **standardised Tsurgeon syntax**
- Rule sets can easily be changed/adapted/extended
- Reordering is invoked only if the reordering rules are given

Implementation details

- Parser runs as a web server
  ⇒ **makes the preordering process independent of the used parser**
- Tsurgeon is modified to enable sentence-wise processing and to avoid infinite loops during tree modifications
  ⇒ **ensures that no SL sentences are lost during preordering**
- Processing parallelised with `GNU parallel` [Tange, 2011]
  ⇒ **leads to a substantially lower reordering time**
  (e.g., reordering of 5,000 segments with the Stanford Shift-Reduce (SR) parser run on 8 cores → the serial implementation took 263.24s, whereas the parallel implementation took 46.10s (5.7 times faster!))

Outline

1. Word Order in SMT
2. Our Approach
3. Experiments and Evaluation
   - SMT
   - NMT
4. Summary and Future Work

Training data (legal domain)

	train	tune	test
EN→DE	1,018,738	500	500
EN→JA	213,592	500	500
EN→ZH	387,275	500	500

SMT models

- Trained with Moses
- Tuned on 500 in-domain sentences using MERT
- Distortion limit = 6
- 5-gram LMs trained with the TL of the parallel corpus
- Word alignment: fast_align

Processing time vs. Translation improvement

- Preordering tested using output of **different parsers** with the **same set of reordering rules**

	Baseline		SR			PCFG			BLLIP		
	BLEU	t_t	BLEU	t_r	t_t	BLEU	t_r	t_t	BLEU	t_r	t_t
EN→DE	40.10	187	40.74	97	254	41.17	372	579	41.49	1279	1468
EN→JA	49.44	135	51.33	25	155	50.29	413	544	51.33	372	492
EN→ZH (PP-NP)	24.99	197	24.40	50	245	24.47	252	460	24.66	627	819
EN→ZH (ofPP)			25.09	49	240	25.22	269	464	25.05	633	820

⇒ **BLEU improves for all parsers!**

- Additional processing time when the fastest parser (SR) is used: 36% for EN→DE, 15% for EN→JA, 22-24% for EN-ZH
- **Positive user feedback justifies longer processing time!**

Training data (legal domain)

	train	tune	test
EN→DE	1,018,738	500	500
EN→JA	213,592	500	500
EN→ZH	387,275	500	500

NMT models

- Trained with OpenNMT
- BPE for EN→DE
- Character-based segmentation for EN→JA/ZH
- Training time: max of 15 epochs

- Automatic (BLEU) & human evaluation (A/B testing) of the NMT outputs:

	Baseline		SR	
	BLEU	Human	BLEU	Human
EN-DE	38.26	49.2	36.74	50.8
EN-JA	67.66	–	60.77	–
EN-ZH (PP-NP)	27.65	36.9	26.67	30.7
EN-ZH (ofPP)			28.75	32.4

⇒ Preordering hurts NMT!

⇒ Confirmed both by automatic, as well as human evaluation

- **Baseline generates the DE verbs in correct positions** (see, e.g., [Bentivogli et al., 2016])
- Preordering seems to have impact both on word order, as well as word choice, e.g.:

EN	The Commission **may**, in any case, **withdraw** such products or substances in accordance with Article37(2).
ENr	The Commission **may**, in any case, such products or substances in accordance with Article37(2) **withdraw**.
B R	Die Kommission **kann** in jedem Fall diese Produkte oder Stoffe gemäß Artikel37 Absatz2 **zurückziehen**. Die Kommission **kann** in jedem Fall solche Erzeugnisse oder Stoffe gemäß Artikel37 Absatz2 **zurückziehen**.
REF	Kommission **kann** in jedem Fall solche Erzeugnisse oder Stoffe gemäß Artikel37 Absatz2 **zurückziehen**.

⇒ Further investigation of preordering for NMT is needed

Summary and Future Work

We presented...

- a fast and customisable preordering component[1]
  for EN$\rightarrow$DE/JA/ZH
- positive impact of preordering on translation quality for SMT
- negative impact of preordering on translation quality for NMT

[1] A simplified version of the preordering component is freely available for research purposes: https://github.com/KantanLabs/KantanPreorder.

We presented...

- a fast and customisable preordering component[1] for EN→DE/JA/ZH
- positive impact of preordering on translation quality for SMT
- negative impact of preordering on translation quality for NMT

Future work for SMT

- Parallelisation with BLLIP parser
- Use of domain-specific parsers to ensure sufficient parsing quality

[1] A simplified version of the preordering component is freely available for research purposes: https://github.com/KantanLabs/KantanPreorder.

We presented...

- a fast and customisable preordering component[1] for EN→DE/JA/ZH
- positive impact of preordering on translation quality for SMT
- negative impact of preordering on translation quality for NMT

Future work for SMT

- Parallelisation with BLLIP parser
- Use of domain-specific parsers to ensure sufficient parsing quality

Future work for NMT

- Further investigation of the impact of the preordering on NMT
- Multi-source approach to combine reordered and non-reordered SL inputs

[1] A simplified version of the preordering component is freely available for research purposes: https://github.com/KantanLabs/KantanPreorder.

Thank you!

Questions?

We thank the EAMT for funding this work!

References I

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo and Marcello Federico:
Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of EMNLP, Austin, USA, 2016.*

Arianna Bisazza and Marcello Federico:
A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena. In *Computational linguistics 42(2), 2016.*

Anita Gojun and Alexander Fraser:
Determing the placement of German verbs in English–to–German SMT. In *Proceedings of EACL, Avignon, France, 2012.*

Roger Levy and Galen Andrew:
Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC, Genoa, Italy, 2006.*

Young-Suk Lee,Bing Zhao and Xiaoqiang Luo:
Constituent Reordering and Syntax Models for English– to–Japanese Statistical Machine Translation. In *Proceedings of COLING, Beijing, China, 2010.*

Ole Tange:
GNU Parallel: The Command-Line Power Tool. In *;login: The USENIX Magazine, February, 2007*.

Chao Wang, Michael Collins and Philipp Koehn:
Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings on EMNLP, Prague, Czech Republic, 2007*.

Peiyu Wu:
Word order errors in Simplified Chinese MT. In *MultiLingual, October/November, 2016*.