

English—Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009

Rejwanul Haque Sandipan Dandapat Ankit Kumar Srivastava
Sudip Kumar Naskar Andy Way

CNGL, School of Computing, Dublin City University, Dublin, Ireland

Introduction

This research represents English—Hindi transliteration in the NEWS 2009 Machine Transliteration Shared Task adding source context modelling into state-of-the-art log-linear phrase-based statistical machine translation (PB-SMT). The Source context model (Stroppa et al., 2007) enables us to exploit source similarity in addition to target similarity, as modelled by the language model. We use a memory-based classification framework that enables efficient estimation of these features while avoiding *data sparseness* problems. We carried out experiments both at character and transliteration unit (TU) level. Position-dependent source context features produce significant improvements in terms of all evaluation metrics.

Context Informed Features

Context Information

$$CI = \{f_{i_k-l} \dots f_{i_k-1}, f_{j_k+1} \dots f_{j_k+l}\}$$

We modify the standard phrase-extraction method of (Koehn et al., 2003) to extract the context information of the source phrases while extracting phrase pairs (\hat{e}_k, \hat{f}_k) .

Context-Informed Features

$$h_{mbl} = \log P(\hat{e}_k | \hat{f}_k, CI(\hat{f}_k))$$

Context-informed features are expressed as the conditional probability of the target phrase \hat{e}_k given the source phrase \hat{f}_k and its context information.

- To avoid sparseness problems, the probability is estimated using *Tilburg Memory Based Learner (TIMBL)* which includes three different classifiers: **IGTree**, **IB1** and **TRIBL** (Daelemans and Antal van den Bosch, 2005).
- Derived Memory-based features h_{mbl} is directly integrated in the state-of-the-art log-linear PB-SMT framework of **Moses** (Koehn et al., 2007).
- Feature weights are optimized using Minimum Error Rate Training (Och, 2003).

An Example: t-table

Source Phrase	Target Phrase	$P(\hat{e}_k \hat{f}_k)$
f1	e1	0.002
f1	e2	0.811
f1	e3	0.021
f1	e4	0.106
f1	e5	0.003
f1	e6	0.001
f1	e7	0.007
f1	e8	0.051

Context-Informed t-table

Source Phrase	Target Phrase	$P(\hat{e}_k \hat{f}_k, CI(\hat{f}_k))$
f1+CI1	e3	0.200
	e5	0.790
	e8	0.010
f1+CI2	e2	0.200
	e4	0.750
	e7	0.040
...	e8	0.010

Results and Conclusion

	S/B	C/TU	Context	ACC
Moses (Baseline)	S	C	0	0.290
		TU	0	0.391
	B	C	0	0.352
		TU	0	0.407
IB1	S	C	± 1	0.391
			± 2	0.386
		TU	± 1	0.406
			± 2	0.359
	B	C	± 1	0.431
			± 2 (NSD1)	0.420
		TU	± 1	0.437
			± 2	0.427
IGTree	S	C	± 1	0.372
			± 2	0.371
		TU	± 1	0.412
			± 2	0.416
	B	C	± 1	0.413
			± 2 (NSD2)	0.407
		TU	± 1	0.445
			± 2	0.427
TRIBL	S	C	± 1	0.382
			± 2 (SD)	0.399
		TU	± 1	0.408
			± 2	0.395
	B	C	± 1	0.439
			± 2 (NSD3)	0.421
		TU	± 1	0.444
			± 2	0.439
S*	C	± 2 (NSD4)	0.419	

Table1: Experimental Results (S/B → Standard / Big data, S* → TM on Standard data, but LM on Big data, C/TU → Character / TU level, SD → Standard submission, NSD → Non-standard submission)

Analysis

- ✓ 10,000 NEs from the NEWS 2009 English—Hindi training data for the standard submissions.
- ✓ Additional English—Hindi parallel person name data (105,905 distinct name pairs) of the Election Commission of India (<http://www.eci.gov.in/DevForum/Fullname.asp>) for the non-standard submissions.
- ✓ In addition to the baseline Moses system, we carried out three different sets of experiments on **IGTree**, **IB1** and **TRIBL**.
- ✓ Each of these experiments was carried out on
 - both **standard data** and **combined larger data**,
 - both at character level (**C-L**) and the TU level (**TU-L**)
 - and considering ± 1 and ± 2 tokens as context (Haque et al., 2009).

Acknowledgements

We would like to thank Antal van den Bosch for his input on the use of memory-based classifiers. We are grateful to SFI (<http://www.sfi.ie>) for generously sponsoring this research under grant 07/CE/11142.

Conclusions

- ✓ Successfully integrated source-context modelling into state-of-the-art PBSMT for English—Hindi Transliteration Task.
- ✓ The accuracy of the **TU-L** and **C-L** baseline systems are **0.391** and **0.290** respectively on **standard datasets**.
- ✓ Furthermore, source-context modelling gives an accuracy of **0.416** and **0.399** for the **TU-L** and **C-L** systems respectively.
- ✓ However, the highest accuracy (**0.445**) was achieved with the **TU-L** system using the **larger dataset**.
- ✓ Source-Context modelling improves accuracy by **43.44%** on the **standard dataset** and by **26.42%** on the **larger dataset** over the Moses *baseline*.
- ✓ **IGTREE** performs better in the **TU-L** system while **TRIBL** seems to perform better in the **C-L** system.

References

- Nicolas Stroppa, Antal van den Bosch and Andy Way. 2007. Exploiting Source Similarity for SMT using Context-Informed Features. *Proceedings of TMI-2007*, Skövde, Sweden, pp. 231-240.
- Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. *Proceedings of ACL-2007*, Prague, Czech Republic, pp. 177-180.
- Rejwanul Haque, Sudip Kumar Naskar, Yanjun Ma and Andy Way. 2009. Using Supertags as Source Language Context in SMT. *Proceedings of EAMT-09*, Barcelona, Spain, pp. 234-241.
- Franz Josef Och, Minimum error rate training in statistical machine translation. 2003. *Proceedings of ACL-2003*, Sapporo, Japan, pp. 160-167.
- Walter Daelemans and Antal van den Bosch. 2005. Memory-based language processing. Cambridge, UK, Cambridge University Press.