TÜBİTAK-UEKAE Statistical Machine Translation System

Coşkun Mermer, Hamza Kaya, Mehmet Uğur Doğan

National Research Institute of Electronics and Cryptology (UEKAE), TÜBİTAK, Türkiye

Dec. 1th, 2009 IWSLT, Tokyo

Introduction

In this study, the TÜBİTAK-UEKAE statistical machine translation system based on the open-source phrase-based statistical machine translation software, Moses, with added components to address the rich morphology of the source languages is presented. Additionally, 3 submissions (primary, contrasive 1, contrasive 2) which use

- linguistic morphological analysis and statistical disambiguation to generate morpheme-based translation models,
- unsupervised subword segmentation to generate non-linguistic subword-based translation models,
- and word-based models but makes use of lexical approximation to cope with out-of-vocabulary words, respectively.

We describe the preprocessing and postprocessing steps and our training and decoding procedures.

Coping with Turkish Morphology

- Turkish is an agglutinative language where words can carry several morphemes in the form of suffixes. *Eg.* Morphological decomposition of the Turkish word and the morpheme-based alignment to its English translation

yap	+a	+ma	+yacak	+sa	+n
do	*be abe to*	*not*	*will*	*if*	*you*

'if you will not be able to do'

- Statistical machine translation involving Turkish requires special attention to Turkish morphology
- Three approaches to dealing with the morphology of Turkish are investigated

Morphological Analyzer

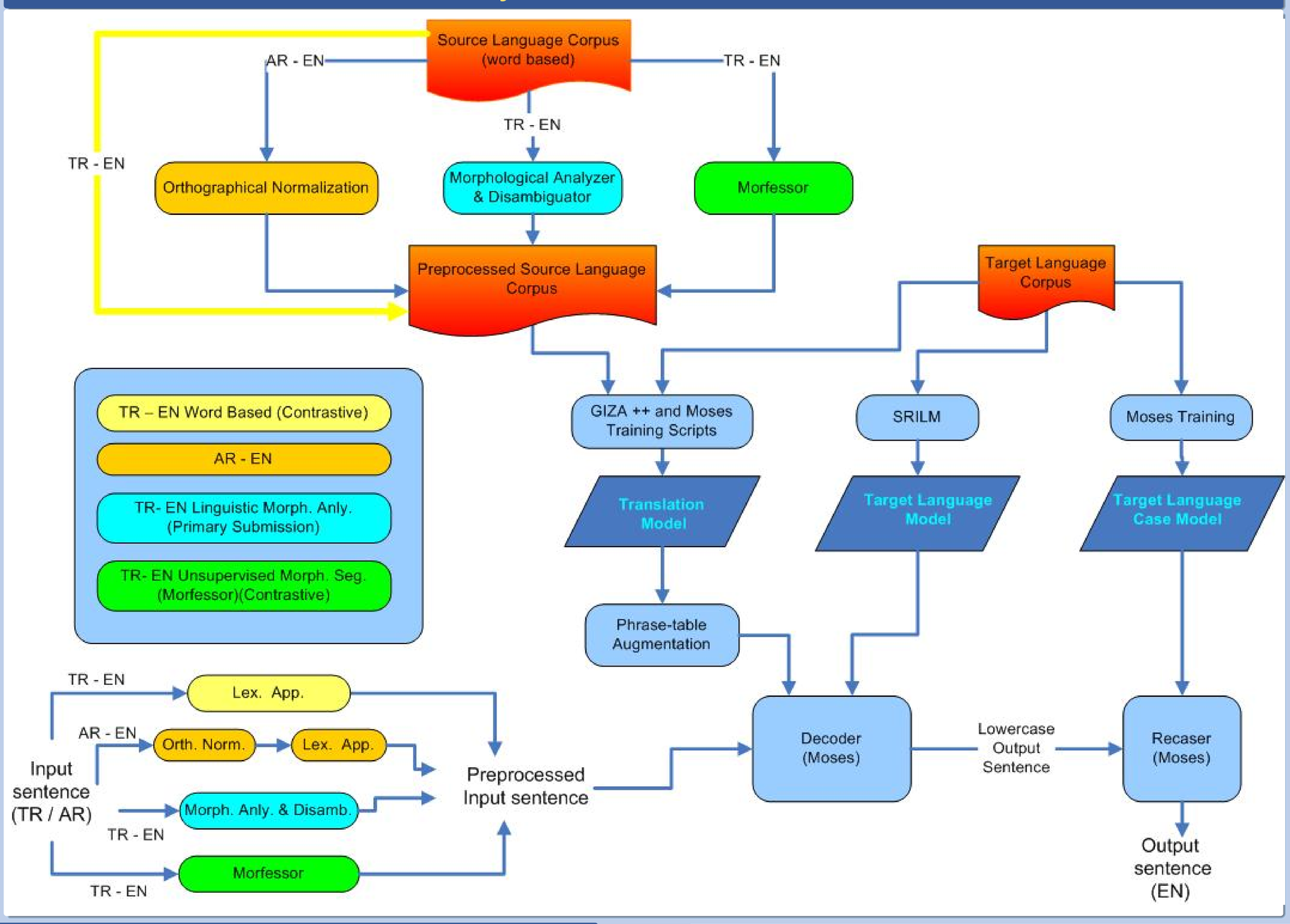
- To separate the words into roots and morphemes linguistic morphological analysis is applied.
- Finite-state morphological analyzer by Kemal Oflazer and statistical disambiguator of Sak et al. are used
- The morphological analyzer output are post-processed to selectively merge or delete some morphemes. *Eg.* The accusative and the imperative markers were deleted from the Turkish corpus; the type-3 infinitive and the "as-if" markers were attached to their roots

Deletion and Attachment of Morphemes

root	morpheme	root	morpheme	root	morpheme	root	morpheme
o	+nu	git	+NULL	yavaş	+ça	dal	+ış
it	*ACC.*	*go*	*IMP. (sing.)*	*slow*	*AS-IF*	*to dive*	*INF-3*
				'slowly'		*'diving'*	
bu adres	+i	dön	+NULL	dikkatli	+ce	uç	+uş
this address	*ACC.*	*turn*	*IMP. (sing.)*	*careful*	*AS-IF*	*to fly*	*INF-3*
				'carefully'		*'flight'*	
ekmek	+i	gir	+in	hızlı	+ca	sat	+ış
(the) bread	*ACC.*	*input*	*IMP. (pl.)*	*quick*	*AS-IF*	*to sell*	*INF-3*
				'quickly'		*'sale'*	
bu düğme	+yi	çağır	+ın	sıkı	+ca	bin	+iş
this button	*ACC.*	*call*	*IMP. (pl.)*	*tight*	*AS-IF*	*to board*	*INF-3*
				'tightly'		*'boarding'*	
a) Deleted Turkish morphemes				*b) Turkish morphemes attached back*			

- Decision to leave a morpheme as a separate unit, to merge with the previous morpheme, or to delete was made based on bilingual human judgments so as to match the English units (i.e., words) as good as possible

IWSLT 2009 - TUBITAK UEKAE System.



Unsupervised Morphological Segmentation

- Development of a morphological analyzer requires lots of manual work and linguistic expertise.
- An unsupervised morphological analyzer, called Morfessor is used.
- Morfessor uses the minimum description length (MDL) principle to find an optimal subword segmentation of a given corpus in the form of a root-and-morpheme vocabulary.
- The segmentations in this model are static in that all the occurrences of a word are assumed to be segmented in the same manner regardless of the context.

Word	Count	Morfessor's segmentation		
anladı	1	anladı		
		understood		
anladım	13	anladım		
		I understood		
anladın	3	anladı	+n	
		understood	*you (sing.)*	
anladınız	1	anladı	+nız	
		understood	*you (plu.)*	
anladıysam	1	anladı	+ysa	+m
		understood	*if*	*I*

Lexical Approximation

- As an alternative to morphological segmentation, we investigated the usefulness of the lexical approximation approach we had previously used in IWSLT 2007
- The corpus and the translation models remain word-based; however, a morphological analyzer may be utilized internally to compute a similarity feature between words based on their shared roots and morphemes

Common System Features

- We used the open-source statistical machine translation toolkit Moses for training the translation models and for decoding.
- An N-gram English language model was trained using the SRI language modeling toolkit
- All the system training and decoding was performed on lowercased and punctuation-tokenized data.
- Although we used 3-gram target language models in our systems
- Table below shows the effect of N-gram model order on the performance of our primary submission.

LM Order	Arabic-English			Turkish-English		
	Dev6	Dev7	*Test 2009*	Dev1	Dev2	*Test 2009*
3	**49.61**	50.52	*49.33*	62.59	59.86	*55.82*
4	49.50	50.91	**50.38**	63.31	**60.33**	**57.24**
5	49.60	**51.18**	*50.34*	**63.48**	60.27	*56.90*

- Similar to our 2007 and 2008 systems, we made use of phrase table augmentation.
- For source vocabulary words that are not included in the phrase table as a result of the phrase extraction process, this technique adds single-word phrase pairs derived from GIZA++produced lexical alignments to the phrase table.
- For some words, forcing the model to propose hypotheses as such may have the negative effect of generating incorrect translations in the output that could have been remedied by other methods (e.g., by lexical approximation)
- Among the provided development corpora, the two most recent sets were reserved for tuning the parameters and internal testing (devsets 1-2 for Turkish and devsets 6-7 for Arabic).
- The remaining corpora were used in training. Hence, for the Arabic-to-English system, devsets1-3 were also added with their 16 references as a training parallel corpus.

Results

case+punc	bleu	meteor	f1	prec	recall	wer	per	ter	gtm	nist
primary	**0.5582**	**0.8120**	**0.8328**	0.8396	**0.8262**	**0.3267**	**0.2676**	**25.219**	**0.7792**	**8.6018**
contrastive1	0.5112	0.7500	0.8008	**0.8529**	0.7547	0.3737	0.3204	28.985	0.7317	6.8455
contrastive2	0.5345	0.7647	0.8015	0.8312	0.7737	0.3486	0.2989	27.611	0.7496	7.6529
no_case+no_punc	**bleu**	**meteor**	**f1**	**prec**	**recall**	**wer**	**per**	**ter**	**gtm**	**nist**
primary	**0.5385**	**0.7763**	**0.8008**	0.8122	**0.7897**	**0.3721**	**0.2932**	**29.029**	**0.7649**	**9.0226**
contrastive1	0.4927	0.7028	0.7573	**0.8200**	0.7035	0.4335	0.3585	33.444	0.7105	6.7275
contrastive2	0.5132	0.7256	0.7659	0.8023	0.7326	0.4026	0.3368	31.872	0.7238	7.6772

- Among the three morphological approaches for Turkish, using morphological analysis customized to the translation task performed the best (primary submission)
- The word-based lexical approximation approach performed close to unsupervised segmentation, even though it was outperformed during development experiments
- In our experiments, training a segmentation model for the English side and using it in system training did not provide a clear improvement over leaving the English corpus as words
- The added complexity of generating roots and morphemes at the decoder output, and the errors in English morphological segmentation could be reasons