FBK @ IWSLT 2010

Arianna Bisazza, Ioannis Klasinas⋆*, Mauro Cettolo, Marcello Federico*

FBK - Fondazione Bruno Kessler
Via Sommarive 18, 38123 Povo (TN), Italy
{bisazza,cettolo,federico}@fbk.eu

⋆IRIT - Université Paul Sabatier
118 Route de Narbonne, F-31062, Toulouse CEDEX 9, France
klasinas@irit.fr

Abstract

This year FBK took part in the BTEC translation task, with source languages Arabic and Turkish and target language English, and in the new TALK task, source English and target French. We worked in the framework of phrase-based statistical machine translation aiming to improve coverage of models in presence of rich morphology, on one side, and to make better use of available resources through data selection techniques. New morphological segmentation rules were developed for Turkish-English. The combination of several Turkish segmentation schemes into a lattice input led to an improvement wrt to last year. The use of additional training data was explored for Arabic-English, while on the English to French task improvement was achieved over a strong baseline by automatically selecting relevant and high quality data from the available training corpora.

1. BTEC task

Turkish and Arabic are morphologically rich languages. When dealing with a small scale task such as the BTEC, this characteristic can have a particularly negative impact on phrase-based statistical MT methods [1]. Following last year's findings [2] we decided to continue working on the problem of out-of-vocabulary words (OOVs) using different strategies. In the Arabic-English pair we tested the usefulness of additional resources by decoding with multiple phrase-tables. As for Turkish-English, we enriched our morphological segmentation rule set and combined several segmentation schemes inside a word lattice. In order to further improve the coverage of the models on the test, we then tried to refine the lexical approximation technique developed last year.

1.1. Arabic-English

The experience of last year taught us that the OOV rate in Arabic-English is indeed a critical issue. This problem is usually addressed by morphological segmentation of the Arabic text prior to training and test. Table 1 shows how the application of the popular segmenter AMIRA [3] affects the BTEC corpora statistics. It can be seen that, despite a considerable reduction of the OOV rate in dev7 from 5.86% to 4.10%, the number of unknown words is still high.

Table 1: *Effect of AMIRA segmentation on BTEC corpora statistics.*

		Training: train+dev{2,3,6}		Test: dev7	
	tokeniz.	$\lvert W\rvert$	$\lvert V\rvert$	$\lvert W\rvert$	OOV
AR	basic	168431	18081	3224	5.86%
	amira	186640	14569	3538	4.10%
EN	*basic*	*193668*	*8517*	*3685*	–

In order to understand if this problem could be overcome by simply adding more training data, we prepared a contrastive run using two translation tables: the first obtained from the BTEC data, the second from a subset of the NIST-MT09 Evaluation data consisting of 186K sentences, 6.2M words (English side) of newswire parallel text. Multiple phrase-table decoding was handled by the Moses decoder [9] in the *'either'* mode, that is for each phrase the union of translation options coming from all the tables is considered.

1.2. Turkish-English

Turkish morphology is agglutinative, which implies that the vocabulary is built by a wide range of basic suffix combinations. Thus it often occurs that a Turkish word is aligned with an English phrase, and sometimes even to a whole sentence as in the following example:

oda	'room'
*oda**m***	'**my** room'
*odam**da***	'**in** my room'
*odamda**yım***	'**I am** in my room'

Previous work [4] has shown that selectively splitting and removing suffixes from the Turkish text used to train a phrase-base SMT system considerably boosts performances in a small scale task like the BTEC. The best segmentation scheme reported in that paper (MS11) mainly includes rules for nominal case and possessive suffixes, plus a few rules on verbal suffixation, namely the splitting of the copula and of the person subject suffixes.

In order to better address the rich verbal morphology we added new segmentation rules for verbs. Note that each of the following schemes includes the rules of the previous:

- **negation (MS13)**: after applying MS11, the suffix *-mA* is extracted from the verb and put after it;
- **ability (MS14)**: the suffix *-Abil*, roughly equivalent to the English modal 'can', is extracted from the verb and put right after it;
- **voice suffixes (MS15)**: passive and causative suffixes are extracted from the verb and put right after it.

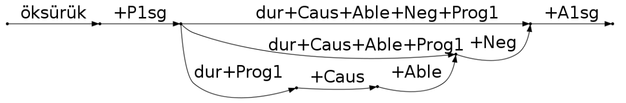
Whenever a verb carries all these suffixes, the application of the last rule set will result in the appearance of 3 new tokens (or 4 if passive and causative are both present) after the verbal root in reversed order.

Table 2: *Application of different segmentation schemes to a Turkish sentence.*

TR:	öksürüğümü durduramıyorum
morph:	öksürük+P1sg+Acc dur+Caus+Able+Neg+Prog1+A1sg
MS11:	öksürük +P1sg dur+Caus+Able+Neg+Prog1 +A1sg
MS13:	öksürük +P1sg dur+Caus+Able+Prog1 **+Neg** +A1sg
MS14:	öksürük +P1sg dur+Caus+Prog1 **+Able** +Neg +A1sg
MS15:	öksürük +P1sg dur+Prog1 **+Caus** +Able +Neg +A1sg
EN:	I can't stop coughing (*litt.* I cannot make my cough stop)

Table 2[1] illustrates the segmentation process: the Turkish text is first morphologically analysed and disambiguated ([5], [6]) and the surface form of suffixes replaced by tags as explained in [4]. The rules for suffix splitting or removal are then applied according to the selected segmentation scheme. It can be seen that in some cases the new rules allow for a better correspondence at the level of words between the Turkish sentence and its English translation. However, this doesn't always corresponds to an improvement in translation quality (see Table 3).

It was shown in [7] that the choice of the optimal segmentation scheme for Arabic-English SMT is not a trivial problem and may depend on several factors such as the training data size. Later [8] obtained considerable gains in translation quality by combining unsegmented and segmented Arabic test sentences into a lattice. Given these findings and given that the segmentation space of Turkish is even richer than the Arabic one, we apply a similar technique: for each sentence of the test set, we combine the outputs of our best rule sets into a segmentation lattice, as shown in Figure 1. In this way the decoder is able to choose the segmentation path that is optimal at the level of words. Edges pertaining to different segmentation schemes are assigned different transition probabilities. For example if 3 schemes are combined the weights are equal to $e^0 = 1$ for the least segmented input, $e^{-0.5} = 0.6065$ for the medium one and $e^{-1} = 0.3679$ for the most segmented one.



TR: öksürüğümü durduramıyorum
EN: I cannot stop coughing (litt. I cannot make my cough stop)

Figure 1: *Segmentation lattice combining different segmentations (MS11, MS13 and MS15) of a Turkish sentence.*

The translation scores obtained on the development set (Table 3) show that the lattice technique performs better than all the simple segmentation schemes tested.

Table 3: *%BLEU–NIST scores obtained on dev2 with different segmentation schemes and with a lattice combination.*

segmentation	BLEU – NIST
MS11	60.30 – 9.367
MS13	58.98 – 9.357
MS14	57.76 – 9.373
MS15	60.32 – 9.575
MS11+13+15	**60.41 – 9.650**

1.3. Evaluation results and discussion

All our systems were built upon the open-source MT toolkit Moses and weights optimized by means of a minimum error training procedure [10].

Arabic-English. Our primary submission is a standard system including AMIRA [3] morphological segmentation of the Arabic text and a distortion limit of 6 words. In the multiple phrase-tables setting each table has its own set of weights, optimized all together by minimum error training. The addition of a translation table trained on out-of-domain data (*contrastive* run) yields a positive reduction of the OOV rate on dev7 from 4.10% to 2.71%. Still, it is not clear whether this is beneficial in terms of translation quality: a consistent gain in terms of BLEU and NIST is reported only on the official test, but not on dev7 and test2009 (see Table 4).

Turkish-English. Here, a similar configuration was used, but the distortion limit was set to 10 to enable the long re-

[1] *öksürük*: 'cough', P1sg: 1st person singular possessive, Acc: accusative, *dur-*: 'stop', Caus: causative, Able: ability, Neg: negation, Prog1: present progressive, A1sg: 1st person singular subject suffix.

Table 4: *%BLEU–NIST scores of the Arabic-English systems on development and test sets.*

Run	ph.tables	dev7	test2009	test2010
primary	btec	**55.02 – 8.735**	**52.04** – 7.494	43.07 – 7.254
contrastive	btec+news	54.20 – 8.620	51.08 – **7.514**	**43.76 – 7.255**

Table 5: *%BLEU–NIST scores of the Turkish-English systems on development and test sets.*

Run	morph.segment.	lex.appr.	dev2	test2009	test2010
primary	MS11+13+15	-	**60.41 – 9.650**	57.70 – 8.612	**53.29 – 8.443**
contrastive1	MS15	-	60.32 – 9.575	**58.28 – 8.660**	52.46 – 8.441
contrastive2	MS11	+	59.68 – 9.513	57.11 – 8.560	51.76 – 8.205
contrastive3	MS11	-	60.30 – 9.367	57.21 – 8.422	52.14 – 8.136

orderings typically required by this language pair. Our primary system (Table 5) was trained on a concatenation of three differently segmented versions of the training corpus (MS11, MS13 and MS15) and tested on a triple segmentation lattice input. The standard set of weights was optimized on MS15-preprocessed data, while the lattice feature weight was directly estimated over a suitable interval by running the decoder several times on the development set. The resulting optimal weight was 0.3. As contrastive runs, we submitted the systems trained and tested on data preprocessed with our best segmentation schemes, that is MS15 and MS11 (*contrastive1* and *contrastive3* respectively). Finally, in the *contrastive2* submission we tried to improve last year's technique of lexical approximation for OOV words: whenever possible, unknown words were replaced by the 3 most morphologically similar words found in the training dictionary. A confusion network was used to provide the options to the decoder, but still no gain in translation quality was observed. Lattice input thus yields the best performance on dev2 and on the official test, whereas MS15 beats it on test2009.

1.4. Conclusions

Working with SMT on morphologically rich languages is a challenge from several points of view. Especially when the training data is limited, statistical approaches suffer from data sparseness, which is only partially leveraged by specific linguistic preprocessing (e.g. morphological segmentation). While the addition of an out-domain translation table didn't yield the expected improvement in Arabic-English, the refinement of the Turkish segmentation scheme and the combination of several schemes into a lattice did benefit the Turkish-English system, showing that there is more to go in this direction. As suggested by previous work on segmented/unsegmented input combination [8] and on selective segmentation (i.e. only infrequent words are decomposed, [11]), it would probably help to apply a similar technique to Arabic.

2. TALK task

2.1. Task description

The TALK task is a new part of the IWSLT competition. The challenge is to translate talks from the TED website[2] from English into French. Talks involve a variety of topics, like photography, psychology, climate change; as a result, it is not possible to classify them in a common category. All considered talks were given in English and transcribed and translated into French by volunteers. In addition to a relatively small parallel corpus of talks, four corpora of various genres were available for the task, whose statistics are summarized in Table 6. The corpus *ALLflt* corresponds to the total available parallel data after filtering out long sentences. The TALK task addressed the automatic translation of both manual and automatic transcriptions, the latter produced by a speech recognition system. Details about these data are also in Table 6.

Table 6: *TALK task corpora statistics.*

Corpus	Sentences	Tokens	
		EN	FR
TED	84k	0.85M	0.89M
NEWS	84k	2.23M	2.54M
EURO	1.68M	46M	50M
UN	7.23M	208M	235M
GIGA	22.5M	663M	773M
ALL	31.5M	920M	1062M
ALLflt	24.6M	558M	633M
DEV	1307	10947	10556
DEV(ASR)	259	11150	10556
TST	3502	26789	27307
TST(ASR)	758	27432	27307

[2] http://www.ted.com/talks/

2.2. Combining additional resources

The significant difference in size of in- and out-of-domain corpora, makes their combination quite difficult. Simply training on all the available data is likely a suboptimal solution, since the contribution of the in-domain data will be too small. The BLEU scores on the dev set for a variety of systems are provided in Table 7. The first system is trained exclusively on the in-domain corpus. The use of all the available data for training the LM leads to a 1.4 BLEU points increase. The linear interpolation of LMs built independently on each corpus results in marginal improvement, while their log-linear interpolation gives an additional 0.5 BLEU point.

On the other hand, training the translation model on all the available corpora results in a 28.61 BLEU score, almost 2 points over using just the in-domain corpus. The linear interpolation of per-corpus built translation tables does not provide any benefit. In this case, it is not advisable to train one model for each available corpus, since the higher number of features seems to cause problems in tuning and decoding. Instead, two models are used, one trained on the in-domain corpus and one on the concatenation of out-of-domain corpora. This system achieves a BLEU score of 29.51 on the dev set, 5 points BLEU more than the system trained only on in-domain data.

Table 7: *%BLEU score on dev set (tokenized, case insensitive). LI stands for "linear interpolation", LLI stands for "log-linear interpolation", PPsel for selection based on perplexity and FRAG for "fragments".*

SYSTEM		BLEU
TM	LM	
TED	TED	24.44
TED	ALL	26.05
TED	LI	26.14
TED	LLI	26.65
ALLflt	LLI	28.61
LI	LLI	28.68
TED, ALLflt	LLI	29.51
TED, FRAG	LLI	29.03
TED, ALLflt	TED, PPsel	29.75
TED, PPsel	TED, PPsel	29.92

2.3. Data selection

In the previous section, the various knowledge sources (corpora) are combined by considering each of them as a whole. On the other hand, it is reasonable to assume that some parts of a corpus are more important than others, both with respect to the considered domain and quality of the contained translations. This is especially true for big corpora crawled from the Web, like the Giga corpus.

To address this issue, a perplexity-based data-selection criterion has been designed: using the in-domain LMs (on source and target languages), all sentence pairs of each out-of-domain corpus are ranked according to their perplexity. Initial experiments showed that the best performance was achieved when the perplexity is computed on the target side only. The amount of selected data was decided empirically by varying the percentage of kept target sentences. Hence, different LMs were trained and used in decoding while the translation table was fixed and trained only on the in-domain data. For the sake of comparison, LMs trained on random samples of the corpus were also tested. Results are plotted in Figure 2. Selection based on the perplexity rank shows indeed to be definitely more effective than random sampling. Reasonably, the difference is much more evident when smaller subsets of the full corpus are used, while for bigger sizes both methods converge to the same point, which coincides with the entry TM=TED and LM=ALL (26.05) of Table 7. The plot also shows validates the assumption that not all sentences are equally important. The best result achieved with perplexity ranking is when 30%[3] of the available corpus is employed. The corresponding scores for the systems using this corpus are given in the lowest block of Table 7. For both language and translation modeling, two distinct models have been trained, one on the in-domain data and one on data selected with the perplexity criterion. If only the LM is replaced, a modest gain is achieved over the system using all the data, that is from 29.51 to 29.75. When the replacement also regards the out-of-domain translation model, an additional gain in BLEU score is achieved, reaching 29.92.

It should be noted that these scores are obtained using only a small fraction of the available data. For language modeling, the training corpus is reduced to 30% of the whole (303M tokens out of 1062M), while for translation modeling optimal results correspond to 48% of the total (303M out of 633M).

2.4. Fragments

The procedure for data selection described above selects data according to its relevance with respect to the task domain. In addition, we also tried to apply a translation quality filter that extracts reliable parallel portions from the selected text pairs. The technique, which is explained in detail in the companion paper [12], has been developed to extracts parallel fragments from comparable documents, that is documents written in different languages and describing the same content, but which are not necessarily direct translations of each other. The technique aims at discovering portions of text that are mutual translations to some extent. We trained our fragment extraction model on the in-domain data and applied it to extract parallel fragments from the Giga data selected through the perplexity criterion. Performance in terms of BLEU score are reported in Table 7, in the last row of the block of results involving the log linear combination (LLI) of five LMs.

[3] While BLEU score is the same also for 20%, NIST score is better with 30% of data

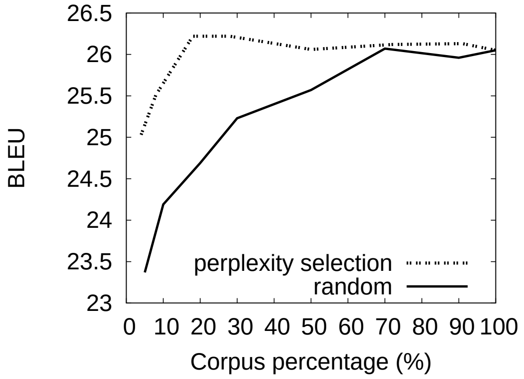


Figure 2: *Dev set %BLEU score with in-domain translation model and varying LM training corpus.*

In particular, the addition of a translation model trained on parallel fragments allows for an improvement of 2.38 BLEU points (from 26.65 to 29.03). This favorably compares with the performance obtained by using all the Giga data (29.51), given that the extracted fragments just account for only 15% of the total data.

2.5. Evaluation results and discussion

All the systems developed uses the Moses decoder. Language models are trained with the IRSTLM [13] language model toolkit, while GIZA++ [14] is used for word alignment. System optimization is obtained by running the MERT [10] procedure on the dev set with respect to the BLEU score. Word reordering has been modeled by means of the lexicalized model available in Moses and limited to 6 positions. In decoding, cube pruning [15] has been activated.

Concerning text pre/post-processing, the typical filtering of long sentences has been applied only to the Giga corpus when employed as parallel resource for the estimation of the translation model. Training texts have been tokenized and lowercased. The automatic translations are de-tokenized and re-cased using tools provided with Moses. Since each caption does not correspond to one full sentence, the first letter is not always capitalized. For proper case restoration, the final punctuation mark of each caption is taken into account to decide on uppercasing or not the first character of the following one. The results of post-processing are shown in Table 8. The fact that case restoration results in a 1.5 BLEU score loss suggests that it might be beneficial to test a true-casing approach instead.

Training on the concatenation of captions that form full sentences resulted in small gains when using only the in-domain data; on bigger systems however there was no improvement and this preprocessing step was omitted in the final setup.

Table 8: *Postprocessing impact on dev set %BLEU score.*

Text format	Dev
tokenized, case insensitive	29.98
tokenized, case sensitive	28.47
detokenized, case sentsitive	27.07

For the official submission, models have been trained on the in-domain corpus and the corpus selected by the perplexity-based scheme. As a result, during decoding two translation tables and two language models are used. In contrast to the previous systems, here cube pruning is not employed, as we discovered later that its omission gives small improvement on the development set. The contrastive system uses all the available data to train the background models. The same systems were used to translate the 1best ASR output. The official scores computed on our submissions are summarized in Table 9. While performance are good on the correct transcription experiments, they are less impressive on the ASR output. This might be due to (i) speech recognition errors and (ii) missing punctuation. If punctuation is not taken into account when computing the BLEU score, the reference transcription score drops from 29.90 to 27.84 while for the ASR condition it increases from 15.19 to 18.35. This means that in the reference transcription task, punctuation is correctly predicted and its omission hurts results. On the other hand, in the ASR translation it is not present and results in a 3 BLEU point loss.

2.6. Conclusions

Summing up, the data selection technique has been proven to be effective, especially on the reference transcription input conditions. We have demonstrated that it is possible to utilize one third or even less of the training data available, and be able to translate better than a system trained on all the corpora available. On the other hand, ASR output scores are quite low. It would be highly recommended to try to translate a richer representation of the ASR decoder, perhaps in the form of confusion networks. In this case, it would also be quite straightforward to insert punctuation information before translating. Another problem is that both systems suffer a 1.5 BLEU score decrease due to incorrectly predicted case information. Maybe training a true-case system could help as well. Finally, an additional issue that could be explored in the future is topic adaptation. In particular, the description available for each talk could be exploited to adapt both the translation and language models of the system on a per-talk basis.

Table 9: *TALK task: official FBK scores.*

	SYSTEM	Dev set		Test set	
		BLEU	TER	BLEU	TER
Reference transcription	Primary	27.07	0.5732	29.90	0.5350
	Contrastive	26.65	0.5781	28.67	0.5436
ASR	Primary	13.18	0.7386	15.19	0.6980
	Contrastive	13.19	0.7403	14.66	0.7022

3. Acknowledgements

This work was supported by the EuroMatrixPlus project (IST-231720), which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development, and by the French Allocation de Recherche contract 26076-2007.

4. References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of HLT-NAACL*, Edmonton, Canada, 2003, pp. 127–133. [Online]. Available: http://aclweb.org/anthology-new/N/N03/N03-1017.pdf

[2] N. Bertoldi, A. Bisazza, M. Cettolo, M. Federico, and G. Sanchis-Trilles, "FBK @ IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009. [Online]. Available: mastarpj.nict.go.jp/IWSLT2009/proceedings/EC_4_fbk.pdf

[3] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in *Proc. of HLT-NAACL: Short Papers*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 149–152.

[4] A. Bisazza and M. Federico, "Morphological pre-processing for turkish to english statistical machine translation," in *Proc. of IWSLT*, Tokyo, Japan, 2009. [Online]. Available: mastarpj.nict.go.jp/IWSLT2009/proceedings/TP_1_bisazza.pdf

[5] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, pp. 137–148, 1994.

[6] T. G. H. Sak and M. Saraçlar, "Morphological disambiguation of Turkish text with perceptron algorithm," in *Proc. of CICLing*, 2007, pp. 107–118.

[7] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proc. of NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, June 2006, pp. 49–52. [Online]. Available: http://www.aclweb.org/anthology/N/N06/N06-2013

[8] C. Dyer, S. Muresan, and P. Resnik, "Generalizing word lattice translation," in *Proc. of ACL: HLT*. Columbus, Ohio, June 2008, pp. 1012–1020. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1115

[9] P. 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, "Moses: Open source toolkit for statistical machine translation," in *Proc. of ACL, Companion Volume: Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180. [Online]. Available: http://aclweb.org/anthology-new/P/P07/P07-2045.pdf

[10] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of ACL*, E. Hinrichs and D. Roth, Eds., 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021.pdf

[11] W. Shen, B. Delaney, A. R. Aminzadeh, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2009 System," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 71–78.

[12] M. Cettolo, M. Federico, and N. Bertoldi, "Mining parallel fragments from comparable texts," in *Proc. of IWSLT*, Paris, France, 2010.

[13] M. Federico, N. Bertoldi, and M. Cettolo, "Irstlm: an open source toolkit for handling large scale language models," in *Proc. of Interspeech*, Melbourne, Australia, 2008, pp. 1618–1621.

[14] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[15] L. Huang and D. Chiang, "Forest rescoring: Faster decoding with integrated language models," in *Proc. of ACL*. Prague, Czech Republic, June 2007, pp. 144–151. [Online]. Available: http://www.aclweb.org/anthology/P07-1019