FBK @ IWSLT 2009

Nicola Bertoldi, Arianna Bisazza, Mauro Cettolo,
Germán Sanchis-Trilles[†]*, Marcello Federico.*

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

[†] Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n, 46022, Valencia, Spain
gsanchis@dsic.upv.es

Abstract

This paper reports on the participation of FBK at the IWSLT 2009 Evaluation. This year we worked on the Arabic-English and Turkish-English BTEC tasks with a special effort on linguistic preprocessing techniques involving morphological segmentation. In addition, we investigated the adaptation problem in the development of systems for the Chinese-English and English-Chinese challenge tasks; in particular, we explored different ways for clustering training data into topic or dialog-specific subsets: by producing (and combining) smaller but more focused models, we intended to make better use of the available training data, with the ultimate purpose of improving translation quality.

1. Introduction

FBK submitted runs at the IWSLT 2009 Evaluation for the Arabic-English and Turkish-English BTEC tasks, and for the Challenge Task involving Chinese and English languages in both directions. This paper reports on efforts we made in the development of such MT systems.

Concerning the Arabic-English and Turkish-English BTEC tasks, a special effort was spent on linguistic preprocessing of the morphologically rich source languages. In particular, we investigated word segmentation techniques which allow for a considerable reduction of the training dictionary and lower the out-of-vocabulary rate of the test set. Moreover, through segmentation, we somehow attacked the problem of mismatch between word formation mechanisms of Arabic and Turkish languages on one side, and English on the other.

In the framework of the Chinese-English and English-Chinese challenge tasks, which involve cross-lingual dialogs, we focused on the language model adaptation problem. Mixtures of n-gram language models are investigated, which are obtained by clustering bilingual training data according to available human annotations. For the sake of adaptation, mixture weight estimation is performed either at the level of single source sentence or test set. Estimated weights are then transferred to the target language mixture model.

The paper is organized as follows. Section 2 describes the linguistic preprocessing techniques applied to Turkish and Arabic languages. In Section 3, the method for LM adaptation applied to the challenge tasks is introduced. In Section 4, the systems employed in the evaluation campaign are sketched and results on development and official evaluation sets presented and discussed. A summary and a list of related issues we will investigate in the next future end the paper.

2. Linguistic Pre-Processing for Morphologically Rich Languages

Indeed linguistic preprocessing plays a fundamental role in any NLP application involving morphologically rich languages, such as Arabic and Turkish. This is particularly true for SMT into English where differences in word granularity between languages reflects on much higher data sparseness on the source side and on the difficulty to properly model word-level alignments. We approached these problems through morphological segmentation of the source languages, referring partly to the work of [1] on an English-Turkish task, and partly to the one of [2] on an Arabic-English task. Secondly, as this was shown to have a positive effect on some Arabic-English SMT systems of previous IWSLT editions [3, 4], we developed two simple language-specific techniques of lexical approximation, which consists in replacing the OOVs of the test set by words of the training that are morphologically close to them.

2.1. Turkish

Turkish is an agglutinative language whose vocabulary is built by a wide range of basic suffix combinations. A Turkish word can thus correspond to a single English word, up to phrases of various length, or even to a whole sentence. On the phonological level, vowel harmony and other phoneme

alternation phenomena systematically lead stems and suffixes to have several surface forms – i.e. allomorphy. Reordering between Turkish and English is also a problematic issue as word alignments are far from being monotonic.

2.1.1. Morphological Analysis

We implemented for Turkish a preprocessing workflow making use of some publicly available linguistic resources and designed from scratch several segmentation schemes that we contrastively tested on the IWSLT09's BTEC task by application on both training and development data.

Our preprocessing workflow starts with morphological analysis, which consists in running K. Oflazer's [5] suffix combinatory FSTs to each entry of the corpus dictionary. This operation is carried out through the *lookup* command of the Xerox Finite-State Tool's suite [6]. As more than one analysis is often possible, disambiguation is performed on the words in context through the perceptron-based tool developed by [7]. As a result of this process, each token is replaced by its lemma followed by a sequence of tags representing lexical features of the analyzed word. The use of feature tags provides a means to abstract from suffix allomorphy.

2.1.2. Segmentation Schemes

The schemes we developed are different combinations of rules determining the splitting or removal of tags from the analyzed words. So far we mainly focused on nominal suffixation and also defined a few rules for the segmentation of verb forms. In order to find an effective rule set we tested eleven morphological segmentation schemes[1] among which *MS11* gave us the best results in term of BLEU scores on the development set. This scheme implies that:

- nominal cases expected to have an English counterpart are split off from words: these are namely dative, ablative, locative and instrumental, often aligning with the English prepositions 'to', 'from', 'in' and 'with/by', respectively. The remaining case tags – nominative, accusative and genitive – are instead removed as they are not expected to have English counterparts;
- possessive tags of all persons are separated from nouns, except the 3rd singular (*P3sg*), which is indeed removed. The tag meaning absence of possessive suffixes is also removed;
- copula is split off;
- person suffixes are split off from finite verb forms and from copula.

The following example shows an analyzed Turkish word before and after segmentation. The number of tokens increases from 1 to 5 as the word is split into noun, possessive, instrumental case, copula and verbal person:

arkadaşımlayım ('I'm with my friend'):

```
arkadaş+Noun+A3sg
   +P1sg
      +Ins
         ^DB+Verb+Zero+Pres
           +A1sg
```

[1] Refer to [8] for a more detailed and linguistically motivated description.

2.1.3. Lexical Approximation

Lexical approximation of Turkish OOVs is performed on the segmented test set and exploits the information produced by the morphological analyzer. The words seen in the training that share the lemma with a given OOV are candidates to its replacement. We designed a simple similarity function that penalizes candidates whose tag sequence differs more from that of the original OOV word, and gives priority to those who share with it a larger number of contiguous tags[2].

Table 1 shows a subset of candidates to the replacement of OOV word *çıkışlar* ('exits', 'checkouts') as ranked by our similarity function. The best result of lexical approximation in this case is the singular form *çıkış* ('exit').

Table 1: *Example of lexical approximation.*

Word	Gloss	Preprocessed (MS11)	Score
çıkışlar	exits	çık+Verb+Posˆ DB+Noun+Inf3+A3pl	
çıkış	exit	çık+Verb+Posˆ DB+Noun+Inf3+A3sg	**93**
çıkma	going out	çık+Verb+Posˆ DB+Noun+Inf2+A3sg	66
çıkacak	will go out	çık+Verb+Posˆ DB+Noun+FutPart+A3sg	66
çıkan	who goes out	çık+Verb+Posˆ DB+Adj+PresPart	44
çıkıyor	is going out	çık+Verb+Pos+Prog1	27
çıkmıyor	isn't going out	çık+Verb+Neg+Prog1	0
çıkarır	takes out	çık+Verbˆ DB+Verb+Caus+Pos+Aor	-15

Words whose lemma was never found in the training remain OOV. Note that in the current implementation the best replacer is chosen in a deterministic fashion before decoding, which raises the chances of introducing noise in the text to translate.

2.2. Arabic

Arabic is also morphologically rich but its segmentation schemes are simpler than those for Turkish, given that the number of involved clitics and suffixes is typically smaller. Nevertheless linguistic preprocessing appears to considerably benefit SMT systems by reducing data sparseness and improving the alignments.

As a prior treatment we perform a specific tokenization (arTok) of Arabic text including removal of short vowels and normalization of extended Arabic Unicode characters and digits.

[2] More precisely: score $=$ match $\times\ 20 - \text{diff}_1 \times 2 - \text{diff}_2 \times 5$, where match, diff_1 and diff_2 are respectively the numbers of shared contiguous tags, different tags in the OOV word, different tags in the replacer candidate.

2.2.1. Morphological Segmentation

Several state-of-the-art tools are available to perform morphological segmentation of Arabic text. For the evaluation we have compared MADA [9] and AMIRA [10], two softwares that differ both on their decision strategy and on the segmentation scheme they apply to the words. While the first is a morphological disambiguator based on linguistic features produced by the Buckwalter analyzer [11], the second is a much lighter-weight SVM classifier based on a -5/+5 character context. As for the segmentation schemes (see Table 2) MADA (scheme "D2") only splits proclitics – namely conjunctions (*w+* 'and', *f+* 'then'), prepositions (*b+* 'by', *k+* 'as', *l+* 'to') and the future tense (*s+*) – whereas AMIRA also separates enclitics, i.e. object and possessive pronouns[3]. Finally, MADA performs orthographic normalization but AMIRA doesn't.

Table 2: *MADA and AMIRA: different segmentation schemes.*

Baseline	wstqwlh			lzmylhA		
	[and-she-will-say-it]			[to-her-colleague]		
	'and she will say it to her colleague'					
MADA	w+	s+	tqwlh	l+	zmylhA	
	[and]	[will]	[she-say-it]	[to]	[her-colleague]	
AMIRA	w+	stqwl	+h	l+	zmyl	+hA
	[and]	[she-will-say]	[it]	[to]	[colleague]	[her]

2.2.2. Lexical Approximation

Lexical approximation of Arabic OOVs is performed on the MADA-preprocessed test set by selectively removing proclitics and enclitics from the unseen word: first the article (*Al+* 'the'), then the object and possessive pronouns (*+ny* 'me', *+y* 'my', *+nA* 'us/our' etc.), verbal person prefixes, *tah marbutah* and eventually the beginning *m+* often used to form participles. The process stops as soon as the word thus being reduced is found in the training, then the substitution is applied to the test prior to translation. If no replacer is found, the OOV word is kept as it is. Note that these are simply surface pattern-matching rules, therefore removal of substrings that are not actual clitics may indeed occur.

3. Online Language Model Adaptation for Spoken Dialog Translation

3.1. Model Adaptation

In systems we developed for the Chinese-English (CE) and English-Chinese (EC) challenge tasks (CT), the LM score $p(\mathbf{e})$ is given either by a single LM (`baseline`) or by the linear interpolation (mixture) of LMs:

$$p(\mathbf{e}) = \sum_{i=1}^{M} w_i p_i(\mathbf{e})$$

where p_i's are target LMs built on clusters which the training data are split in. With the help of Figure 1, the basic adaptation procedure is described in the following.

Let us assume that the parallel training data have been partitioned into a set of `M` bilingual clusters, according to some criterion. On each cluster, language specific LMs are estimated, which are then organized into two language specific mixture models. All operations described so far are performed off-line. Now let us consider a source text or sentence to be translated. Before translation, the input is used to estimate optimal weights of the source language mixture through Expectation-Maximization. The resulting weights are then transferred to the target language mixture, which is finally used as LM feature function by the SMT system.

3.2. Clustering

The manual annotation of IWSLT dialogs is exploited for clustering purposes. In fact, the CT data is provided with dialog annotations, which allows the use of complete dialogs as single units. Each dialog is represented as a bag of both source and target words. The use of texts of both languages was suggested by the slight gain obtained in a preliminary investigation.

For the clustering of dialogs, the CLUTO[4] package was employed. Its setup includes the `direct` clustering algorithm, which computes the *k*-way clustering directly, and the cosine distance as similarity function between dialogs in their array representation. The number of clusters tested is 2, 4, 6 and 8; on each of them a different LM was trained (see Figure 1). Additional LMs were built on the complete BTEC+CT data.

3.3. On-line weight optimization

Once different LMs have been estimated on clusters built from training data, they are interpolated at translation time with weights that need to be estimated. For this purpose, several approaches were investigated, which are described in the following.

3.3.1. Set specific weights

The LM-interpolation weights were estimated on the source side of the complete test set. This approach, which is the most straightforward, has nevertheless an important drawback: the estimated weights are those that well model the whole test set on average, without considering possibly significant differences between specific sentences. Hence, the potential benefit of estimating several LMs may fade.

3.3.2. Sentence specific weights

In this case, one specific set of weights is estimated for each sentence of the test set. By doing so, we expect that the effect of separating the training corpus into several subsets yields

[3] AMIRA splits the same proclitics as MADA except for the future tense.

[4] Available from http://glaros.dtc.umn.edu/gkhome/views/cluto

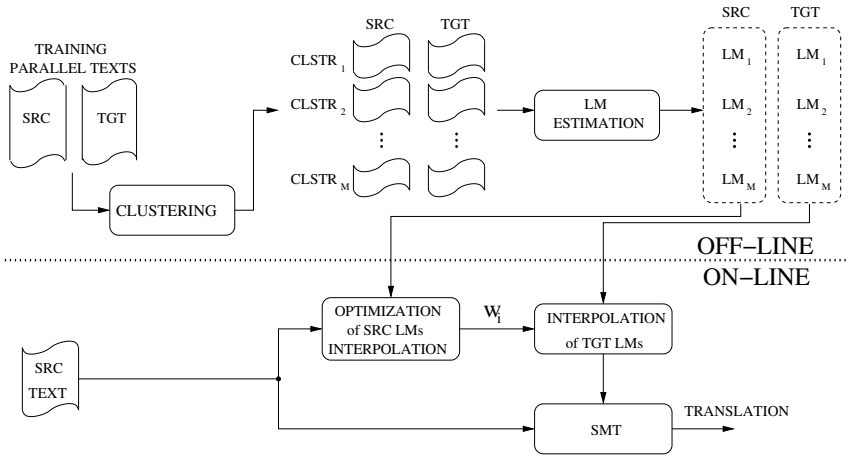

Figure 1: *Basic procedure for LM adaptation.*

better results, since the EM procedure is allowed complete freedom in assigning the LM weights. However, weights computed in such a manner may be less reliable, since the estimation is performed on few data (one single sentence).

3.3.3. Two-step weight estimation

This approach merges the previous two in the attempt of keeping their advantages and overcoming the drawbacks. Once sentence specific weights have been computed, each (source) sentence is assigned to the specific cluster corresponding to the most weighted LM. This being done, one set of weights can be re-estimated for each one of the clusters obtained in this way. This approach has the intuitive benefit of mirroring the clustering of the training data into the test set, while still avoiding the possible data sparseness issue that can affect the sentence specific weight estimation.

4. Evaluation results

4.1. Baseline System

Given a string $\mathbf{f}$ in the source language, the goal of statistical machine translation [12] is to select the most probable string $\mathbf{e}$ in the target language. By assuming a log-linear model [13, 14], the optimal translation can be searched for with the criterion:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \max_{\mathbf{a}} \sum_{r=1}^{R} \lambda_r h_r(\mathbf{e}, \mathbf{f}, \mathbf{a}),$$

where $\mathbf{a}$ represents a word- or phrase-based alignment between $\mathbf{f}$ and $\mathbf{e}$, and $h_r(\mathbf{e}, \mathbf{f}, \mathbf{a})$ $r = 1, \ldots, R$ are *feature functions*, designed to model different aspects of the translation process. In particular,

$$h(\mathbf{e}, \mathbf{f}, \mathbf{a}) = \log p(\mathbf{e})$$

provides the log score of the target LM.

Our systems are built upon the open-source MT toolkit Moses [15]. The decoder features a statistical log-linear model including a phrase-based translation model, a language model, a distortion model and word and phrase penalties. The weights λ_r of the log-linear combination are optimized by means of a minimum error training (MERT) procedure [16]. The phrase-based translation model provides direct and inverted frequency-based and lexical-based probabilities for each phrase pair included in a given phrase table. Phrase pairs are extracted from symmetrized word alignments generated by GIZA++ [17].

4.2. Turkish-English System

4.2.1. Data

For training our Turkish-English system we exclusively used the provided BTEC training corpus. Parameters were tuned on IWSLT09's devset1 using the gold reference translation only. Evaluation during development was performed on devset2.

4.2.2. Baseline Setup

The baseline preprocessing consists in simple tokenization (IWSLT09's released script) and lowercasing of the source side data. Due to the severe mismatch in word order between the languages, we set the distortion limit (DL) to 10. Moses option *-drop-unknown* was active in all submitted runs.

4.2.3. Results

Table 3 shows how morphological segmentation affects the statistics of Turkish texts. First of all, our best segmentation scheme – MS11 – reduced the OOV rate by more than half. In addition, it reduced the differences in token granularity between Turkish and English, by increasing the number of Turkish words in the training corpus (from 6.9 to 8.4 words/sentence on average whereas there are 9.1 on the English side), and lowering the number of different forms. Fi-

nally, a decrease of the test set's cross-entropy[5] estimated through a 5-gram LM trained on the provided Turkish data was observed.

Table 3: *Effect of preprocessing on training corpus size and dictionary (Turkish side), OOV and cross-entropy of devset2.*

	training		devset2	
preprocessing	|W|	|V|	OOV%	H(bits)
-	139,514	17,619	6.16	59,435
MS11	168,135	10,450	2.54	57,379

These positive effects reflect indeed on the translation quality, resulting in a 5 points BLEU improvement over the baseline (Table 4 - primary).

On the other hand, lexical approximation (contrastive1) does not improve BLEU in *-drop-unknown* conditions[6], despite the observed reduction of OOV: from 2.54% to 0.89% in devset2, from 2.14% to 0.95% in the test. This may be due to the noise introduced by the deterministic choice of 1-best OOV replacer.

Given the short average size of IWSLT corpora sentences, we also tested translation performances in unlimited distortion conditions (contrastive2): this results in a slight gain on the official evaluation, whereas DL=10 gave us the best performances on devset2.

Table 4: *Features and %BLEU on devset2 and IWSLT09 BTEC TE test set of baseline and submitted systems.*

system	MS11	lex.appr.	DL	devset2	test
baseline	-	-	10	54.80	N/A
primary	+	-	10	59.77	56.77
contrastive1	+	+	10	59.24	56.75
contrastive2	+	-	∞	59.02	57.04

4.3. Arabic-English System

4.3.1. Data

The Arabic-English system was trained on the provided BTEC training corpus, to which we added devsets 2, 3 and 6 (only with gold reference translation). Minimum error training was run on IWSLT09's devset1 using all references. Evaluation during development was performed on devset7.

[5]We chose cross-entropy to compare LMs across different segmentations schemes as its computation does not involve normalization on the number of tokens. A conventional dictionary upper bound size of 10^7 is assumed to make LMs with different OOV rates more comparable, although care must be taken in interpreting these figures.

[6]When decoding without the *-drop-unknown* option we registered a 0.2 points BLEU absolute improvement.

4.3.2. Baseline Setup

Similar to the Turkish-English baseline, but with default distortion limit (DL=6).

4.3.3. Results

Arabic-specific advanced tokenization (Table 5 - baseline1) alone is responsible for a 0.5 point BLEU improvement over the baseline (baseline0). Morphological segmentation through MADA (primary)[7] yields an additional gain of 2.3 points on devset7. As for AMIRA (contrastive2), results are inconsistent through the test data: this segmentation technique indeed performs 1 point better than MADA on the official test, while we observed the reverse during our development experiments. Indeed it is difficult to draw qualitative conclusions on the tools used, as they differ on several levels. Although AMIRA splits more clitics than MADA does, the two segmenters yield to roughly the same number of OOV words both in devset7 and test.

Lexical approximation results (contrastive1) are also discrepant: while no improvement was possible on devset7, BLEU increases by around 0.7 points on the official test.

The last submitted run (contrastive3) differs from our primary system only by the LM smoothing method. Final evaluation results confirm that improved-kneser-ney (or modified-shift-beta) smoothing performs better than plain kneser-ney, despite some inconsistencies we had observed during development.

Table 5: *Features and %BLEU on devset7 and IWSLT09 BTEC AE test set of baseline and submitted systems.*

system	preproc.	lex.appr.	LM	devset7	test
baseline0	-	-	msb	51.87	N/A
baseline1	arTok	-	msb	52.38	N/A
primary	mada	-	msb	54.68	52.23
contrastive1	mada	+	msb	54.52	52.92
contrastive2	amira	-	msb	54.60	53.36
contrastive3	mada	-	kn	53.78	51.92

4.4. English-Chinese System

4.4.1. Data

For training, both CT and BTEC English-Chinese training corpora were used. Statistics are shown in Table 6.

MERT was run on the development set of the CT task; after that, that set was included into the training corpus. The development sets of previous evaluation campaigns were not included as such into the training data, but only their vocabulary. This in accordance with a preliminary investigation, where we observed that these sets did not improve the quality of the translation of the CT development set.

[7]Both MADA and AMIRA morphological segmentations were performed on arTok-preprocessed data.

Table 6: *Statistics of the EC training data: running words (|W|), vocabulary size (|V|) and average sentence length ($\bar{s}$).*

EC	ENG			CHI		
task	| W |	| V |	$\bar{s}$	| W |	| V |	$\bar{s}$
BTEC	153K	7294	7.7	172K	8428	8.6
CT	119K	3271	11.8	102K	3737	10.2

The same system, only differently tuned, was employed for ASR and CRR conditions: the weights of the log-linear model were estimated on the corresponding versions of the development set.

4.4.2. Baseline Setup

The setup of the baseline system for the EC CT task was the same used for the other language pairs, with distortion limit set to 6.

4.4.3. Results

For establishing the most effective system to be used for the evaluation campaign, experimental investigation was conducted on the CT development set, which was split into two parts, used for tuning and evaluation, respectively. In this stage, the models were trained exclusively on the training data. We compared the use of different numbers of clusters (Section 3.2) and the three on-line weight estimation methods described in Sections 3.3.1 (`set`), 3.3.2 (`sbs`) and 3.3.3 (`2steps`). Results in terms of %BLEU score are reported in Figure 2 which includes as reference line the score of the baseline, that is the system with a single LM.

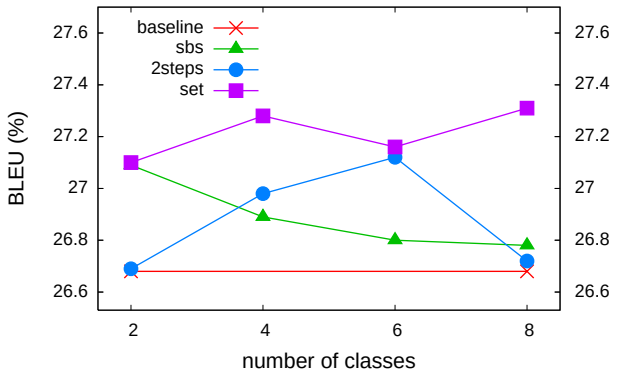

Figure 2: *Performance of the baseline and of the three weight estimation methods as functions of the number of clusters. Scores are computed on a portion of the CT development set.*

First of all, it can be noted that whatever the number of clusters and the scheme followed for the estimation of LM interpolation weights, the proposed adaptation technique yields quite interesting improvements. More specifically, it seems that the `set` estimation method guarantees better performance, but unfortunately here and also in the other plots not reported here its behaviour is quite unstable. On the other side, the shape of the curve of the `2steps` method is - not only here but typically - unimodal, fact that makes its behavior more predictable. Unfortunately, the impressive improvements we measured in terms of perplexity by applying those adaptation techniques were mirrored into translation quality only to a limited extent.

According to those outcomes, we clustered the CT training data into six dialog clusters and decided to submit as primary run the system which dynamically estimates the weight of the linear interpolation of the six LMs via the `2steps` procedure. Table 7 reports results on the official test set of the evaluation campaign of all our submissions, namely the primary, the baseline and the other two contrastive runs corresponding to the other two remaining on-line weight estimation methods.

It can be seen that our adaptation technique clearly performs better in the case of the ASR input. In this case, and in contrast with the Chinese-English direction (see below), it is the `set` setup the one that performs best, and the `2steps` technique yields somewhat mixed results, specially in terms of BLEU.

Table 7: *Official results in terms of %BLEU and precision/recall (p/r) for the EC CT.*

		ASR		CRR	
system	submission	BLEU	p/r	BLEU	p/r
baseline	contrastive3	32.75	61.7/59.1	40.40	68.4/66.3
2steps	primary	33.37	63.2/59.4	40.05	68.5/66.1
set	contrastive1	33.71	63.3/59.6	40.33	68.8/66.2
sbs	contrastive2	33.20	63.2/59.5	39.73	68.1/65.7

4.5. Chinese-English System

4.5.1. Data

The use of the available data for the CE CT task was the same as for the opposite direction. Statistics of the training data are shown in Table 8. Differences with figures reported in Table 6 are due to the fact that, on both directions, in the source side punctuation marks and case information are removed while they are kept in the target side.

Table 8: *Statistics of the CE training data: running words (|W|), vocabulary size (|V|) and average sentence length ($\bar{s}$).*

CE	CHI			ENG		
task	| W |	| V |	$\bar{s}$	| W |	| V |	$\bar{s}$
BTEC	148K	8408	7.4	183K	8344	9.1
CT	89K	3734	8.9	141K	3696	14.0

4.5.2. Baseline Setup

For the CE CT, the setup of the baseline was the same of the opposite direction; of course, specific tuning was performed.

4.5.3. Results

The setup of our adaptation technique was performed in the same way as for the opposite direction; official results are collected in Table 9.

Table 9: *Official results in terms of %BLEU and precision/recall (p/r) for the CE CT.*

		ASR		CRR	
system	submission	BLEU	p/r	BLEU	p/r
baseline	contrastive3	30.01	63.3/63.2	31.82	66.4/67.3
2steps	primary	30.13	63.5/63.4	31.92	66.5/67.8
set	contrastive1	29.92	63.6/62.7	32.15	66.5/67.6
sbs	contrastive2	29.96	64.0/63.6	31.87	66.7/67.6

These results, although quite mixed, show that a slight improvement can be obtained by our adaptation technique, specially in terms of precision and recall (`p/r`). More specifically, the `2steps` technique is the only method able to outperform the baseline in both ASR and CRR conditions.

5. Summary and Future work

The evaluation has shown how specific linguistic preprocessings can benefit a purely statistics-based, language-independent NLP application like SMT. In particular we have proved that selectively splitting/removing suffixes from morphologically analyzed Turkish text considerably improves translation quality.

In the next future we would like to refine our Turkish segmentation schemes by better addressing verbal suffixation. Concerning lexical approximation of both Turkish and Arabic, it may be helpful to feed Moses with multiple options of replacement so that the translation and language models would contribute to the decision at decoding time.

Concerning the LM adaptation method, it allowed a quite limited improvements of the translation quality; anyway, in not reported experiments, we observed impressive gains in terms of perplexity, which proves that there is much room for improvement and hence that the approach deserves to be further investigated. Future work will likely be carried out on larger tasks than BTEC, like Europarl and those of NIST MT evaluation campaigns, and will involve different issues left out from this paper. For example, unsupervised clustering, i.e. grouping sentences without the help of any manual annotation; or the use of development or even test data for guiding the clustering. Another subject to be studied is the re-use of weights estimated for the optimal interpolation of source LMs also for interpolating target LMs; in fact, a source-to-target weight map could be learned from parallel development/training set which is expected to guarantee more effective mixtures of LMs.

6. 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 Spanish MEC under scholarship AP2005-4023 and grant CONSOLIDER Ingenio-2010 CSD2007-00018.

7. References

[1] K. Oflazer and I. D. El-Kahlout, "Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation," in *Proc. of Workshop on SMT*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 25–32.

[2] N. Habash and F. Sadat, "Arabic Preprocessing Schemes for Statistical Machine Translation," in *Proc. of NAACL HLT*. New York City, USA: Association for Computational Linguistics, June 2006, pp. 49–52.

[3] H. K. C. Mermer and M. U. Dogan, "The TUBITAK-UEKAE Statistical Machine Translation System for IWSLT 2007," in *Proc. of IWSLT*, Trento, Italy, 2007, pp. 176–179.

[4] T. A. W. Shen, B. Delaney and R. Slyh, "The MIT-LL/AFRL IWSLT-2008 MT System," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 69–76.

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

[6] K. R. Beesley and L. Karttunen, *Finite State Morphology*. Palo Alto, CA: CSLI Publications, 2003.

[7] T. G. H. Sak and M. Saraclar, "Morphological Disambiguation of Turkish Text with Perceptron Algorithm," in *Proc.of CICLing*, 2007, pp. 107–118.

[8] A. Bisazza and M. Federico, "Morphological Pre-Processing for Turkish to English Statistical Machine Translation," Technical paper at IWSLT09.

[9] N. Habash and O. Rambow, "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," in *Proc. of ACL*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 573–580.

[10] K. H. M. Diab and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in *HLT-NAACL 2004: 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.

[11] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0," Linguistic Data Consortium, University of Pennsylvania, 2002, lDC Catalog No.: LDC2002L49.

[12] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–312, 1993.

[13] A. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[14] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of ACL*, PA, Philadelphia, USA, 2002.

[15] P. Koehn et al., "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proc. of the ACL Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.

[16] F. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of ACL*, Sapporo, Japan, 2003, pp. 160–167.

[17] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.