Online Language Model Adaptation for Spoken Dialog Translation

Germán Sanchis-Trilles[1], *Mauro Cettolo*[2], *Nicola Bertoldi*[2], *Marcello Federico*[2]

(1) Instituto Tecnológico de Informática, Valencia, Spain
(2) FBK - Ricerca Scientifica e Tecnologica, Trento, Italy
gsanchis@dsic.upv.es, {cettolo,bertoldi,federico}@fbk.eu

Abstract

This paper focuses on the problem of language model adaptation in the context of Chinese-English cross-lingual dialogs, as set-up by the challenge task of the IWSLT 2009 Evaluation Campaign. Mixtures of n-gram language models are investigated, which are obtained by clustering bilingual training data according to different available human annotations, respectively, at the dialog level, turn level, and dialog act level. For the latter case, clustering of IWSLT data was in fact induced through a comparable Italian-English parallel corpus provided with dialog act 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. Experimental results show that, by training different specific language models weighted according to the actual input instead of using a single target language model, significant gains in terms of perplexity and BLEU can be achieved.

1. Introduction

This paper focuses on spoken dialog translation, in which specific states of dialog can be identified and exploited to adapt a phrase-based statistical machine translation (SMT) system, in order to compute more appropriate translations. Adaptation in this scenario, more than coping with a statistical mismatch between training and testing data, is considered as a means to introduce additional *context* in the system.

In particular, we augment the target language model (LM) component, by introducing parameters that are adapted on the input text. The LM is implemented as a mixture of sub LMs, that are estimated through some bilingual clustering of the training data.

Experiments have been performed on the IWSLT 2009 Challenge Task, for Chinese and English languages on both directions, with the correct transcription of spoken documents as MT input – i.e. no recognition errors, but without punctuation or case information. Different clusterings have been considered and compared, which take into account different kinds of similarities among the contained sentences. Clustering methods we considered are all based on available human annotations. In particular, sub-LMs have been estimated on the basis of the manual annotation of IWSLT texts, in terms of dialog identifier and speaker role. Moreover, another corpus annotated at the dialog act level has been also exploited to infer a clustering of the IWSLT training data.

The paper is organized as follows. Section 2 briefly lists some of the papers dealing with related issues. Section 3 depicts the adaptation procedure. Data, tested systems and results are presented and discussed in Section 4. Future directions for upgrading the approach are listed in Section 5, while final remarks are given in Section 6.

2. Related work

Adaptation in SMT is a research field that is receiving an increasing amount of attention. Although applying their techniques to interactive MT, one of the first approaches to this task was performed by [1], in which they followed the ideas by [2] and added cache LMs and TMs to their system. Later, [3] developed a system that was able to learn from its own errors. For doing that, they started with a rule-based system, and then trained a SMT system using as source language the output of the rule-based system and as target language the correct reference. In [4], different ways to combine available data belonging to two different sources was explored; in [5] similar experiments were performed, but considering only additional source data. In [6], alignment model mixtures were explored as a way of performing topic-specific adaptation, the alignments being used only to extract phrases.

A work that resembles the one presented here is [7], where each source sentence was used to build a query and retrieve similar sentences from a larger corpus. Then, a specific LM was trained and interpolated with a generic LM. Finally, this combination was used to translate the original sentence. In [8], each sentence was used to select similar data within the same corpus by means of TF-IDF, data with which prepare specific LMs and TMs ready to be interpolated.

3. Model adaptation

Given a string $\mathbf{f}$ in the source language, the goal of statistical machine translation [9] is to select the most probable string $\mathbf{e}$ in the target language. By assuming a log-linear model [10, 11], 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. In our approach, such score 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.

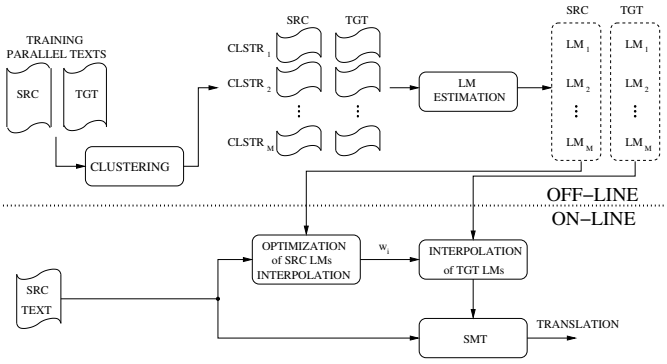

Figure 1: *Basic procedure for LM adaptation.*

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.

4. Experiments

4.1. IWSLT data

The IWSLT 2009[1] evaluation campaign is carried out using BTEC (Basic Travel Expression Corpus), a multilingual speech corpus containing tourism-related sentences. In addition, a collection of human-mediated cross-lingual dialogs in travel situations are provided to the participants of the Challenge Task (CT). We focused on the CT, correct recognition results, Chinese-English (CE) and English-Chinese (EC) language pairs. The CT corpus includes for each sentence a dialog identifier and the speaker class, i.e. agent, customer or interpreter. Table 1 reports statistics (running words and vocabulary size) of the training corpora used in our experiments after the preprocessing performed by means of the tools supplied by the organizers; the numbers for the two directions are different, despite the original texts are the same, because casing and punctuation have been removed from source texts, but kept on target texts.

Table 1: *Statistics of the IWSLT training data. $|W|$ stands for running words, $|V|$ for vocabulary size and $\bar{s}$ for average sentence length.*

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

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

Since we are going to exploit speaker and dialog annotations of the CT corpus, more detailed statistics are reported in Table 2. The figures regard the target side for the CE task; those for the other direction are quite similar and hence omitted for clarity.

Table 2: *Speaker-based statistics of the CT training set.*

	speaker	|W|	|V|	$\bar{s}$
agent	native	46.7K	2240	14.8
	interpreter	26.8K	1626	14.1
customer	native	33.3K	2082	13.9
	interpreter	33.8K	1878	12.9

In order to study the similarity, or better the differences, between training and testing conditions, we computed the same statistics for the CT development data set (Table 3). It clearly results that the two sets differ not only for the sentence length, but also in terms of distribution of utterances from the interpreter. We will see later if and in which cases this mismatch affects the system performance.

Table 3: *Speaker-based statistics of the CT development set.*

	speaker	|W|	|V|	$\bar{s}$
agent	native	2.5K	427	15.1
	interpreter	0.8K	218	13.2
customer	native	0.5K	152	11.8
	interpreter	1.7K	307	12.3

[1] http://mastarpj.nict.go.jp/IWSLT2009/

4.2. Nespole! data

Nespole![2] (NEgotiating through SPOken Language in E-commerce) [12] was a EU funded project, running during years 2000-2002. It aimed at providing a system capable of supporting advanced needs in e-commerce and e-service by resorting to automatic speech-to-speech translation. In particular, one of the two implemented showcases supported multilingual negotiations and discussion between a tourist information/service provider (a so-called destination) and a customer who wanted to organise a trip exploring all available possibilities, including travel, accommodation, attractions and recreation, cultural events, dining and so on. Collected data mirrored such scenario. For the purposes of the work presented in this paper, 58 Nespole! dialogs were used; they were collected in 2000 involving Italian speakers, then translated into English and manually labeled in terms of dialog acts. Table 4 reports corpus statistics regarding the English side of the dialogs, while Table 5 provides the (self-explanatory) labels and counters of the most frequent dialog acts.

Table 4: *English side statistics of the Nespole! dialogs.*

#turns	| W|	|V|	$\bar{s}$
2522	15335	1344	6.1

Table 5: *Most frequent Nespole! dialog acts.*

label	counter
`give-information`	963
`affirm`	408
`descriptive`	285
`request-information`	199
`acknowledge`	122
`greeting`	80
`negate-give-information`	62
`thank`	55
`request-action`	55
...	...
total	2522

Nespole! texts are quite different from IWSLT texts, although both of them are tourism-related. Just think that the cross-corpus perplexity is around 900, while the perplexity of IWSLT development/test sets ranges approximately from 50 to 200. Nevertheless, Nespole! data include valuable semantic annotation which is worth exploiting.

4.3. Baseline system

The baseline system is built upon the open-source MT toolkit Moses [13].[3] 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 of the log-linear combination are optimized by means of the Minimum Error Rate Training (MERT) procedure [14]. Optimization of the weights of the log-linear combination was only performed for the baseline system, and these weights were re-used for all other systems. Although there could be reasons for reestimating this set of weights, we did not do so in order to be able to better isolate the effects of including different LMs. Moreover, we chose not to include each one of different LMs into the MERT optimization so as to ensure its stability.

The phrase-based translation model provides direct and inverted frequency-based and lexical-based probabilities for each phrase pair in the phrase table. Phrase pairs are extracted from symmetrized word alignments generated by GIZA++ [15]. A 5-gram word-based LM is estimated on the target side of the parallel corpora using the improved Kneser-Ney smoothing [16]. For modeling word reordering, in addition to a negative-exponential on the reordering distance, a model conditioned on phrase pairs was estimated, namely the "orientation-bidirectional-fe" distortion model [17].

4.4. Clustering

In this section the use of manual annotation of IWSLT and Nespole! texts for clustering purposes is described. The clusters are the starting point for building adapted SMT systems as described in Section 3.

4.4.1. IWSLT

Two types of clustering have been built by exploiting the annotation of training/development texts (see Section 4.1), one related to dialogs and the other to speakers.

Dialog based clustering: 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 both sides 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.

Speaker (agent/customer/interpreter) based clustering: four clusters are built exploiting this type of annotation, namely one of agent turns, one of customer turns, and two of interpreter turns which are translations of agent and customer utterances, respectively. LMs were then estimated on both sides of each cluster. In this case, additional LMs were trained on BTEC and BTEC+CT data.

[2] http://nespole.itc.it

[3] Available from http://www.statmt.org/moses/

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

4.4.2. Nespole!

The English side of Nespole! data (see Section 4.2) has been employed for clustering the IWSLT training data. Three 5-gram LMs were estimated: two on `give-information` and `request-information` sentences, and one on the remaining texts. The rationale behind this choice is that those two dialog acts are expected to label quite different sentences in terms of lexicon, syntactic structure and punctuation (when available). The three LMs have been successively used for partitioning the English side of the IWSLT training data (BTEC+CT) on the basis of perplexity; the same clusters have been mirrored to the Chinese side. LMs have been finally estimated on such clusters (see upper part of Figure 1) in addition to the LMs estimated on the whole source and target IWSLT texts.

4.5. Online 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.

4.5.1. 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.

4.5.2. 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 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).

4.5.3. 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.

4.5.4. Oracle weight estimation

So as to provide an upper bound of the performance that can be reached with the adaptation technique presented in this paper, optimal sentence specific weights have also been estimated on the reference translations.

4.6. Results

Coherently to what written at the beginning of this section, experiments were performed on the development sets IWSLT09 of the CT, CE/EC, correct recognition result transcripts tasks. They were split in two parts (DEV1 including 4 dialogs, DEV2 with 6 dialogs) which were alternatively used for MERT and evaluation.

Results are provided in Figures 2-5. Each of them includes four plots: the two plots on the left show BLEU scores, those on the right perplexity; the two upper plots refer to results measured on DEV1 with DEV2 used for MERT, vice-versa for the two lower plots. Figures 2 and 4 report results obtained by dynamically estimating the interpolation weights at the sentence level (Section 4.5.2), while Figures 3 and 5 refers to the two-step technique (Section 4.5.3). Finally, Figures 2 and 3 show performance for the EC direction, while Figures 4 and 5 for the CE task.

The five curves in each plot refer to different systems:

`baseline`: SMT system using one single LM estimated on the whole training corpus (Section 4.3);

`dialog`: interpolation of LMs built on the dialog based clustering as described in Section 4.4.1;

`nespole`: interpolation of LMs built on the clustering induced by the Nespole! data (Section 4.4.2);

`ACI`: interpolation of LMs built on the speaker based clustering as described in Section 4.4.1;

`oracle`: the LMs are those built on the dialog basis, but the interpolation weights are estimated as described in Section 4.5.4.

Results achieved by interpolating LMs with weights estimated at the test set level (Section 4.5.1) are not reported for the sake of simplicity and because they are not better than those of the competing techniques, as expected.

Before the detailed analysis, a general comment is that in terms of perplexity the idea of building LMs on some motivated partition of the training data and then interpolate them with weights estimated on the actual input performs very well, yielding significant improvements whatever the clustering technique, the number of clusters (LMs) and the scheme followed for the estimation of interpolation weights. Moreover, the BLEU score of the `oracle` system confirms that the approach is really appealing. On the other side, for the fair systems the impressive improvement in terms of perplexity is not always mirrored in the BLEU score, especially

for clusters built exploiting either Nespole! annotation or speaker information, for which sometimes a degradation is even observed.

In relation to our experimental outcomes, the following additional remarks can be made:

- the `oracle` curves are unimodal with a peak at six clusters, which is then the optimal number of LMs to be interpolated;
- the shape of the curves of the two-step procedure (Figures 3,5), although are not higher than those of the estimation performed on single sentences (Figures 2,4), are more similar to those of the oracle (unimodal), fact that makes its behavior more predictable;
- the `dialog` based clustering improves or at least does not worsen too much baseline BLEU scores, even if it tends to be quite far from the oracle quality; there is no clear evidence about the optimal number of clusters;
- `ACI` works quite well for the EC task but not for the CE direction;
- `nespole` clustering does not seem to be effective in terms of BLEU score;
- performance by switching the role of DEV1 and DEV2 is quite different;
- improvements over the baseline are larger on EC direction than on CE.

It is important to stress the fact that training/development and test conditions were quite different in the experiments conducted. This was already pointed out by the comparison of figures in Tables 2 and 3, but it is even more evident by observing that MERT is effective only for the EC direction and when DEV2 and DEV1 are used for development and evaluation respectively, while it degrades the performance of the initial setup in all the other three cases; Table 6 gathers the variations of the BLEU score between initial and final configurations of the SMT system for the two directions (CE and EC) and with the two possible roles for DEV1 and DEV2. This disappointing behavior is probably due to the too small size of DEV1, fact that could also explain why our adaptation technique does not work very well on DEV2, i.e. when DEV1 is used for development.

Table 6: *MERT effect on the BLEU score.*

test on	mert on	Δ BLEU CE	Δ BLEU EC
DEV1	DEV2	-0.19	+3.39
DEV2	DEV1	-0.67	-1.12

It can also be observed that, in some rare cases, the `oracle` BLEU scores drop below the `dialog` scores. This could be due to the fact that we assume that the interpolation weights computed on the reference sentence are the ones that best exploit the provided models. However, such assumption could not be true in the case of a severe mismatch between such models and reference sentences, leading to the possibility of achieving better scores with other weights.

A final remark is needed on the fluctuating performance of the ACI clustering. Its purpose is to obtain speaker-role specific LMs, which should theoretically perform better than generic LMs when it is possible to know which is the role of the actual speaker. However, if training and test conditions within each dialog role present a severe mismatch, as seems to be the case according to Tables 2 and 3, such an approach is bound to yield a very limited benefit, if any.

Despite all the precautions required by the fact that the experimental outcomes are not unquestionable, an encouraging conclusion can be drawn. It emerges that the LM adaptation approach proposed here is promising and can guarantee quite stable improvements over the baseline quality when the clustering is built at the level of dialogs and the interpolation weights are estimated with the two-step scheme.

5. Future work

One interesting issue left out from this paper is unsupervised clustering. In fact, here LMs have been estimated on partitions of training data built exploiting manual annotations. Data partitioning can also be obtained in an unsupervised manner, by just applying the clustering algorithm to the surface form of the training data. The clusters will be built according to the criterion determined by the similarity function used by the algorithm, that in the simplest case could be lexical based, but that could involve even more sophisticated linguistic knowledge.

A second matter to be examined involves the use of development or even test data for guiding the clustering. In fact, here training data have been partitioned disregarding the actual sentences to be translated. The assumption is that training texts are similar to texts processed at run-time. However, it can happen that they differ, even quite a lot like the BTEC training data and the CT development/test sets in the IWSLT 2009 evaluation campaign. A possible solution to this problem is that of partitioning the development/test data and then induce that clustering to the training data; in such a way, the set of the resulting LMs should be able to better model the variety of input sentences.

Another issue which deserves an investigation regards the interpolation of target LMs by re-using weights estimated for the optimal interpolation of source LMs. In fact, although it appears as a reasonable choice, it could happen that the likelihood on the target side is maximized with different weights than those which ensures the maximum likelihood on the source side. A source-to-target weight map could be learned from a parallel development/training set.

Finally, we are going to assess our techniques on larger tasks than BTEC, like Europarl and those of NIST MT eval-

uation campaigns. This will involve also the schemes that seemed to be ineffective on the IWSLT tasks, since even for them evidence of room for improvement has been found in experiments presented here.

6. Conclusions

This paper has presented a technique for adapting the LM of SMT systems to the actual input. The assumption is that the LM is provided as a linear interpolation of sub-LMs, each estimated on a specific portion of the training data. The interpolation weights are then estimated dynamically on the text to be translated via a maximum likelihood EM-based procedure.

Different methods for partitioning training data and different schemes for weight estimation have been experimentally tested on the data released for the IWSLT 2009 evaluation campaign. Impressive improvements in terms of perplexity have been observed, while the quality of the automatic translation, as measured by the BLEU score, increased in a less evident manner; nevertheless, the dialog-based partitioning and the two-step weight estimation scheme seem to guarantee quite stable improvements over the baseline quality. In any case, much room for improvement has been shown, which suggests that the approach deserves to be further investigated.

7. 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.

8. References

[1] L. Nepveu, G. Lapalme, P. Langlais, and G. Foster, "Adaptive language and translation models for interactive machine translation," in *Proc. of EMNLP*, 2004.

[2] R. Kuhn and R. D. Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[3] M. Simard, C. Goutte, and P. Isabelle, "Statistical phrase-based post-editing," in *Proc. of NAACL HLT*, 2007.

[4] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proc. of ACL WMT*, 2007.

[5] N. Bertoldi and M. Federico, "Domain adaptation in statistical machine translation with monolingual resources," in *Proc. of EACL WMT*, 2009.

[6] J. Civera and A. Juan, "Domain adaptation in statistical machine translation with mixture modelling," in *Proc. of ACL WMT*, 2007.

[7] B. Zhao, M. Eck, and S. Vogel, "Language model adaptation for statistical machine translation with structured query models," in *Proc. of CoLing*, 2004.

[8] Y. Lü, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization," in *Proc. of EMNLP*, 2007.

[9] 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.

[10] 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.

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

[12] A. Lavie, F. Pianesi, and L. S. Levin, "The NESPOLE! System for multilingual speech communication over the Internet," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1664–1673, 2006.

[13] 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. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-2045

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

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

[16] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 4, no. 13, pp. 359–393, 1999.

[17] P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation," in *Proc. of IWSLT*, Pittsburgh, PA, 2005.

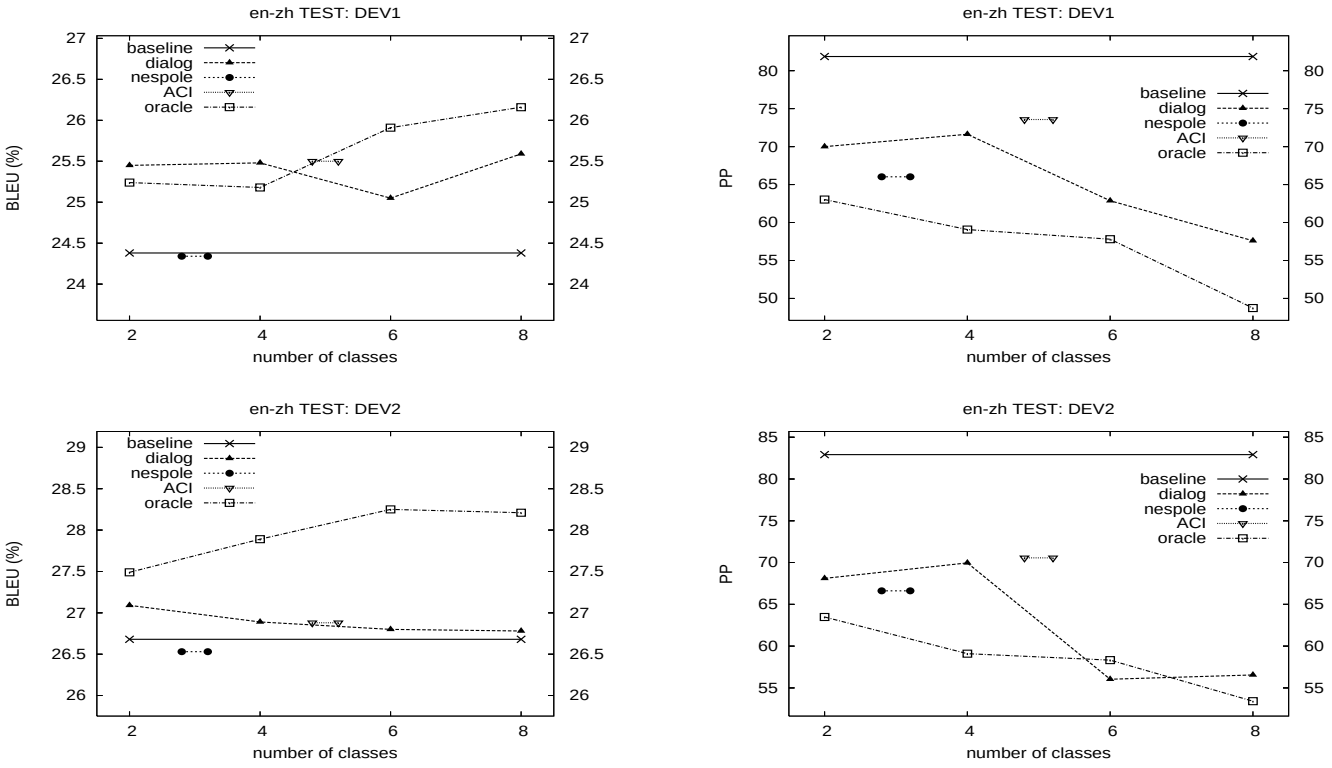

Figure 2: *EC results (BLEU scores and perplexity) with different clustering methods, sentence specific weight estimation.*

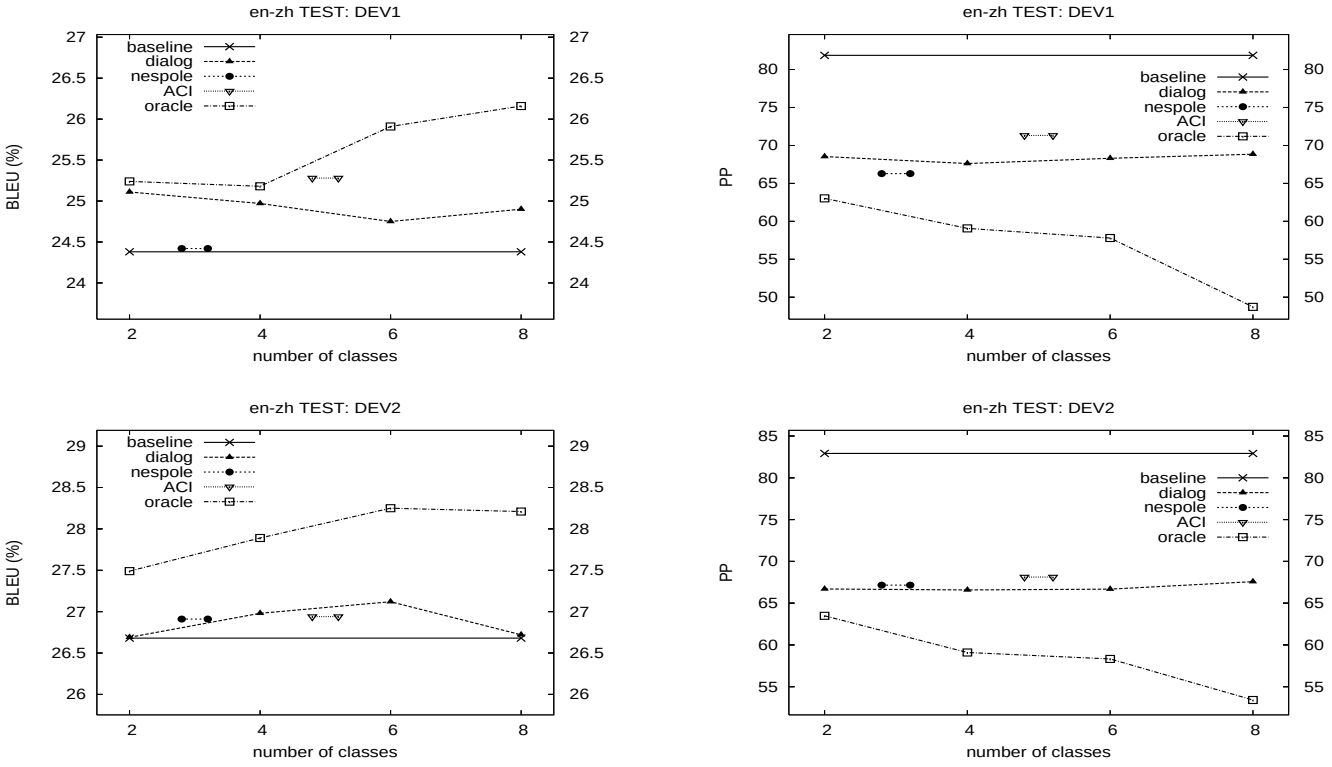

Figure 3: *EC results (BLEU scores and perplexity) with different clustering methods, two-step weight estimation.*

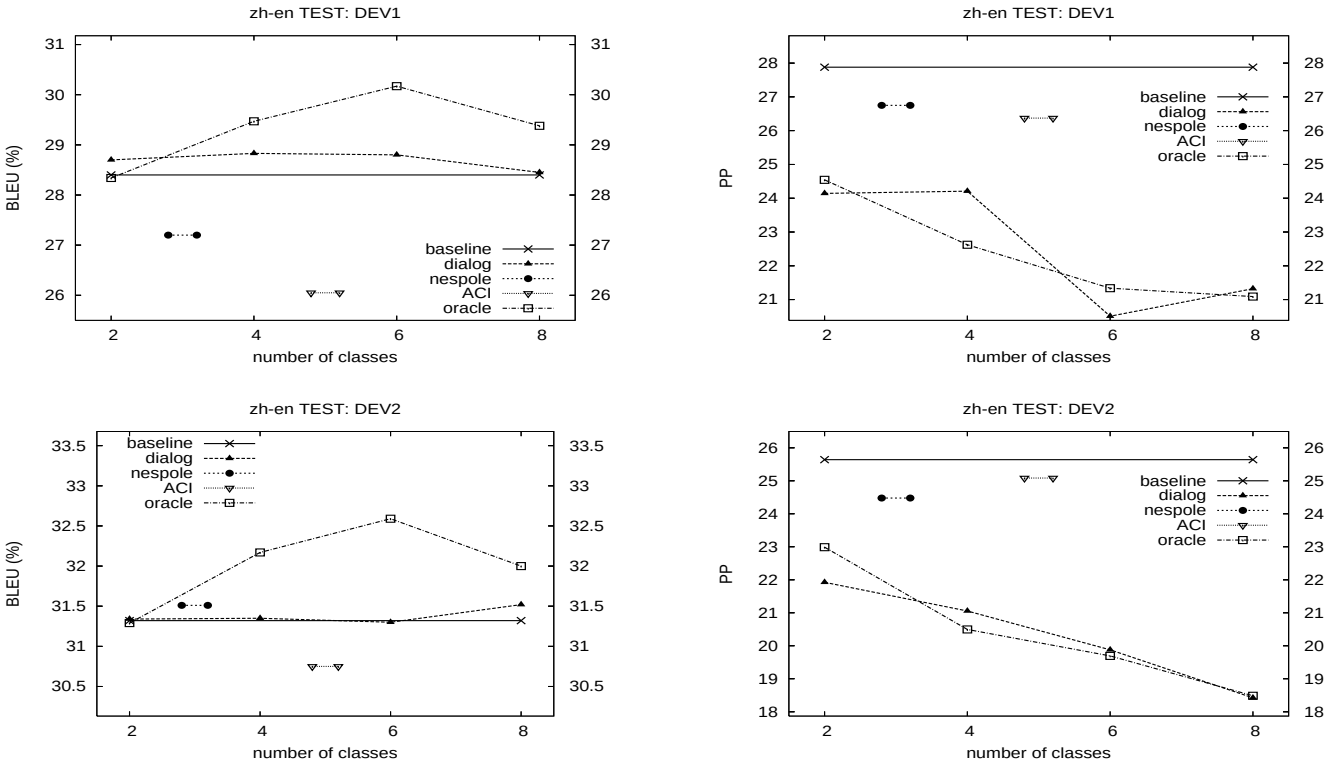

Figure 4: *CE results (BLEU scores and perplexity) with different clustering methods, sentence specific weight estimation.*

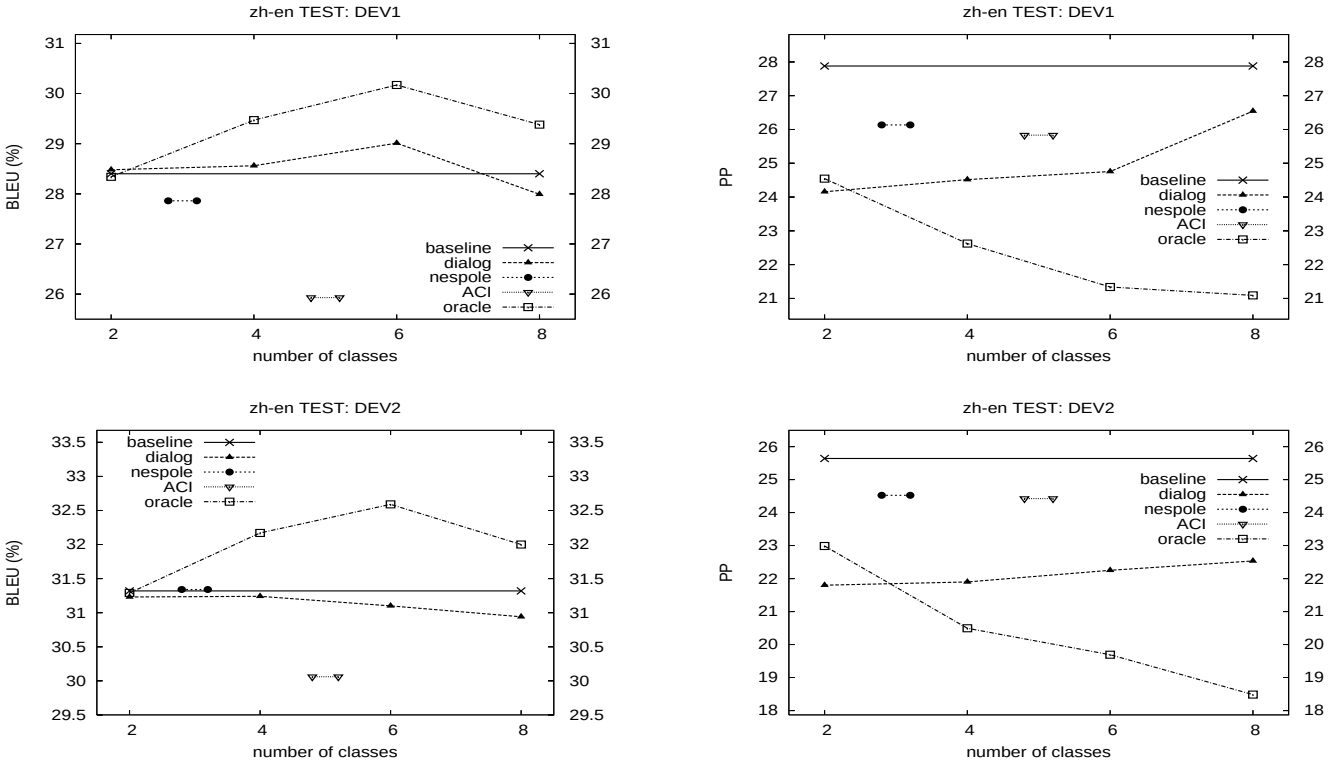

Figure 5: *CE results (BLEU scores and perplexity) with different clustering methods, two-step weight estimation.*