Web crawling and domain adaptation methods for building English–Greek machine translation systems for the culture/tourism domain

Víctor M. Sánchez-Cartagena
Prompsit Language Engineering
Av. Universitat s/n. Edifici Quorum III
E-03202 Elx, Spain
vmsanchez@prompsit.com

V. Papavassiliou S. Sofianopoulos P. Prokopidis
Institute for Language and Speech Processing
Athena Research and Innovation Center
Athens, Greece
{vpapa, s_sofian, prokopis}@ilsp.gr

Abstract

This paper describes the process we followed in order to build English↔Greek machine translation systems for the tourism/culture domain. We experimented with different data sets and domain adaptation methods for statistical machine translation and also built neural machine translation systems. The in-domain data were obtained by means of the ILSP Focused Crawler.

1 Introduction

The total contribution of the tourism and travel industry to Greece GDP in 2014 was over 29 400 million euro and represented 17.3% of its total GDP (Turner, 2015). The availability of web content written in English and related to Greek tourist spots and travel advice is a crucial factor for ensuring a pleasant experience to foreign tourists. Additionally, foreign companies (e.g. airlines, car hire companies, etc.) may also need to translate their content into Greek (for instance, technical documentation for their local workforce). Machine translation (MT) can help to reduce the cost of producing this content.

In this paper, we explore the use of web crawling and domain adaptation methods for building English↔Greek MT systems targeting the Greek tourism and travel domain.[1]

In the remainder of this paper, we describe our web crawling approach, as well as the different domain adaptation methods we have followed for building phrase-based statistical MT (PBMT) systems and some preliminary work on neural MT (NMT). We share our findings about the translation performance of the different systems we evaluated with the aim that it is useful to other MT practitioners. In particular, Section 2 summarises the most relevant related publications, Section 3 lists the corpora we used to train our MT systems (including the ones we crawled from the web), Section 4 describes our experiments with domain adaptation in PBMT, and Section 5 presents the NMT systems we built. The paper ends with some concluding remarks and future work directions.

2 Related work

When building an MT system addressed to a specific domain, one cannot totally discard the training data that is not related to that domain (out-of-domain data), since it can increase the lexical coverage of the system and improve the translation of those expressions that are translated in the same way regardless of the domain, especially if the in-domain data is limited in size. However, out-of-domain data should not negatively affect the translation of domain-specific expressions. The way in which in-domain and out-of-domain data are combined in order to maximise translation quality is usually called *domain adaptation*. Multiple domain adaptation methods can be found in the literature. As testing them all is impractical, our work is focused on those that have been employed regularly in MT shared tasks such as WMT[2] and IWSLT.[3]

Other authors have evaluated different domain

[1] Many Greek tourist spots are related to the ancient Greek civilization: MT systems produced should be able to deal with texts related to archaeology, Greek mythology, etc.

[2] http://www.statmt.org/wmt16/

[3] https://sites.google.com/site/iwsltevaluation2016/

adaptation methods for a particular translation task. The most similar work is that by Pecina et al. (2015), who already experimented with different domain adaptation methods for building English–Greek PBMT systems. However, there are multiple differences between their work and ours: (i) their work addressed the domains of environment and labour legislation, while ours is focused on the tourism domain; (ii) we experiment with multiple domain adaptation methods they did not include in their experimental set-up; and (iii) we explore the feasibility of the emerging NMT paradigm in our work.

Some other relevant works are those by Toral et al. (2016), who target the tourism domain and South-Slavic languages and explore the use of web crawling and phrase table interpolation, and Durrani et al. (2013), who evaluate the impact of phrase table interpolation and data selection on a news translation task for 10 language pairs, to name but a few.

3 Data used

The main resources from which our MT systems were built are a parallel corpus and two monolingual corpora (English and Greek) obtained by crawling content from museum, archaeological and tourism-related websites with the ILSP Focused Crawler (Papavassiliou et al., 2013). We will refer to them as in-domain corpora. We put aside a few websites from which we extracted the test and development sets. 157 websites were crawled in order to obtain the training corpora, while the test and development sets were respectively obtained from 46 and 49 websites.[4]

We also used a set of out-of-domain corpora which are not related to the Greek tourism domain and were combined with the in-domain corpora by means of domain adaptation methods. We used all the freely available English–Greek parallel corpora from the Opus project (Tiedemann, 2012) plus the English–Greek part of the PGV parallel corpus (Prokopidis et al., 2016). Our out-of-domain Greek monolingual set includes corpora obtained by crawling the 2009-2012 online archives of major Greek daily newspapers, plus the Greek side of all parallel corpora. Concerning English, we used monolingual corpora released for the WMT 2016 shared translation task, the British National Corpus[5] and the English side of the parallel corpora. Tables 1 and 2 show the number of sentences and tokens in each corpus. We report total sizes excluding *OpenSubtitles 2016* (OST) since it was not used in some of the experiments described in Section 4.

Corpus	# sentences (k)	# tokens (M) en/el
in-domain		
train	96	2.0/1.9
development	2	0.043/0.042
test	2	0.045/0.044
out-of-domain		
DGT	1 973	45.1/46.3
ECB	96	2.5/2.4
EMEA	362	5.7/6.1
Europarl	1 240	34.3/34.5
GNOME	6	0.05/0.05
JRC-Acquis	12	0.3/0.3
KDE4	117	0.7/0.8
OST 2016*	25 542	241/205
PGV	63	1.3/1.4
SETIMES2	225	5.4/5.7
SPC	8	0.2/0.2
Tatoeba	2	0.02/0.02
Ubuntu	6	0.03/0.03
Wikipedia	99	1.9/1.9
Total w/o OST	4 207	97.6/99.8
Total	29 748	338.8/305.6

Table 1: Parallel corpora

All the corpora were tokenized and truecased.[6] Pairs of sentences longer than 80 tokens on either side were removed from the parallel corpora, as well as duplicated sentences.We evaluated the systems we built in terms on the BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and chrF1 (Popović, 2015) evaluation metrics. Statistical significance of the difference between systems is computed with paired bootstrap resampling (Koehn, 2004) ($p \leq 0.05$, 1 000 iterations).

4 Domain adaptation for statistical machine translation

We built our PBMT systems with Moses (Koehn et al., 2007). The weights in the log-linear model were tuned by means of MIRA (Watanabe et al., 2007) and the in-domain development corpus described in Setion 3. KenLM (Heafield, 2011, default parameters) was used to estimate 5-gram language models (LMs).

We experimented with different domain adaptation methods: for each direction, we identified the

[4]The list of crawled websites is available at: `http://abumatran.eu/tourism-culture.txt`.

[5]`http://www.natcorp.ox.ac.uk/`

[6]The truecaser model was trained from the concatenation of all the available corpora for each language

Corpus	# sentences (k)	# tokens (M)
in-domain English		
crawled	477	10.8
in-domain Greek		
crawled	863	19.0
out-of-domain English		
parallel	29 760	338.7
parallel w/o OST	4 219	97.6
News Commentary	388	9.8
News Crawl	131 084	3 074.1
News Discussions	50 301	917.8
BNC	6 029	108.4
Total w/o OST	192 921	4 232.8
Total	218 463	4 473.9
out-of-domain Greek		
parallel	29 760	305.6
parallel w/o OST	4 219	99.7
newspaper-crawled-data	5,700	142.4
Total w/o OST	9 910	242.1
Total	35 451	448.0

Table 2: Monolingual corpora

best performing translation model (TM) and LM adaptation methods and built the final PBMT system by combining them. We also compared their translation performance with that of the straightforward concatenation of all the corpora.We evaluated the following TM adaptation strategies:[7]

- `Log-linear`: log-linear combination (Koehn and Schroeder, 2007) of a phrase table built from the in-domain corpus and a phrase table built from the concatenation of the out-of-domain parallel corpora (from now on, *out-of-domain phrase table*).
- `Fill-up`: combination by means of *fill-up* (Bisazza et al., 2011) of the in-domain and out-of-domain phrase tables and reordering tables.
- `Linear2`: linear interpolation of the in-domain and out-of-domain phrase tables. Weights were obtained by perplexity minimisation on the development set following the *weighted counts* algorithm by Sennrich (2012).
- `Linear14`: linear interpolation (Sennrich, 2012) of 14 phrase tables: one for each parallel corpus listed in Table 1.
- `Data selection`: linear interpolation (Sennrich, 2012) of the in-domain phrase table and the two phrase tables produced by the data selection strategy followed by Rubino et al. (2014). They applied the algorithm by Axelrod et al. (2011) in order to split the out-of-domain parallel corpora into a *pseudo in-domain* set and a *pseudo out-of-domain* set and filtered the latter by means of vocabulary saturation (Lewis and Eetemadi, 2013).[8]
- `Data selection + fill-up`: as Sennrich (2012) did not define his linear interpolation strategy for reordering tables, we also re-evaluated the most promising approaches that involve linear interpolation of phrase tables after combining the in-domain and out-of-domain reordering tables by means of fill-up (Bisazza et al., 2011).

Concerning the adaptation of the LM, we built an in-domain LM from the in-domain monolingual corpus and an out-of-domain one from the concatenation of all the out-of-domain data and combined them in the ways listed below.[9]

- `Linear interpolation` with weights that minimize perplexity on the TL side of the development set. Weights were obtained by means of SRILM (Stolcke and others, 2002).[10]
- `Log-linear` combination.[11]
- `Data selection`: log-linear combination of the in-domain LM and an LM built from a subset S of the out-of-domain monolingual data obtained by means of data selection. We followed the strategy by Ruiz et al. (2012) in order to select the subset: S contains the top sentences with the minimum cross-entropy difference (Moore and Lewis, 2010) and its size is chosen so as to minimise the perplexity of the TL side of the development set on an LM trained on S.

[7]Unless explicitly mentioned, when we evaluated each strategy, reordering tables and LMs were obtained from the concatenation of all the corpora.

[8]We used the same values as them for the data selection parameters with the exception of the proportion of sentences included in the pseudo-in-domain set. We selected it by means of the same approach we describe for monolingual data selection.

[9]In the evaluation of these strategies, phrase and reordering tables were obtained from the concatenation of all the parallel corpora.

[10]Due to the enormous memory requirements, we could only evaluate this method for the English-to-Greek direction.

[11]`http://www.statmt.org/moses/?n=FactoredTraining.BuildingLanguageModel#ntoc1`

System	BLEU	TER	chrF1
Basic systems			
In-domain data	0.2524	0.6527	0.6564
Out-of-domain data	0.2072	0.7030	0.6232
All data (concatenation)	0.2680	0.6396	0.6666
Translation model adaptation			
Log-linear	0.2563	0.6490	0.6576
Fill-up	0.2701	**0.6370**	**0.6689**
Linear2	0.2703	0.6380	**0.6691**
+ fill-up	**0.2714**	0.6372	**0.6698**
Linear14	0.2699	0.6390	**0.6688**
Data selection w/o OST	**0.2717**	**0.6360**	**0.6695**
+ fill-up	0.2697	**0.6363**	**0.6693**
Data selection w/ OST	0.2686	0.6388	**0.6708**
+ fill-up	0.2702	0.6375	**0.6710**
Language model adaptation			
Log-linear	**0.2765**	**0.6325**	**0.6704**
Linear interpolation	**0.2713**	0.6380	0.6671
Data selection w/o OST	**0.2765**	**0.6325**	**0.6704**
Data selection w/ OST	**0.2752**	**0.6339**	**0.6707**
Best LM + best TM	**0.2758**	**0.6318**	**0.6705**

Table 3: Results of English-to-Greek domain adaptation experiments.

System	BLEU	TER	chrF1
Basic systems			
In-domain data	0.3173	0.5779	0.5129
Out-of-domain data	0.2820	0.6098	0.4770
All data (concatenation)	0.3376	0.5640	0.5460
Translation model adaptation			
Log-linear	0.3194	0.5760	0.5137
Fill-up	0.3385	**0.5587**	0.5474
Linear2	0.3396	**0.5562**	**0.5480**
+ fill-up	**0.3418**	**0.5560**	**0.5491**
Linear14	0.3383	**0.5583**	0.5475
Data selection w/o OST	0.3400	**0.5570**	**0.5480**
+ fill-up	**0.3412**	**0.5562**	**0.5490**
Data selection w/ OST	**0.3453**	**0.5534**	**0.5589**
+ fill-up	**0.3459**	**0.5505**	**0.5596**
Language model adaptation			
Log-linear	**0.3430**	**0.5554**	**0.5509**
Data selection w/o OST	0.3368	**0.5608**	0.5464
Data selection w/ OST	**0.3449**	**0.5578**	**0.5509**
Best LM + best TM	**0.3513**	**0.5486**	**0.5628**

Table 4: Results of Greek-to-English domain adaptation experiments.

We only included the OST parallel corpus in the training corpora for the systems built with data selection methods (we also report translation quality obtained without OST for them) because this corpus contains very informal language that can hardly suit the tourism/travel domain if it is used as a whole, and it is so large that it would produce a slow PBMT system with a big memory footprint unless only a subset is used.

Tables 3 and 4 show the results of the evaluation. Scores of the strategies that perform statistically significantly better than the simple data concatenation are shown in bold. For both language directions, the combination of the best TM and LM domain adaptation strategies (whose names are shown in bold) delivers better translation quality than the data concatenation according to the three evaluation metrics.

The relative performance of each method, however, varies across directions. For English-to-Greek, LM adaptation methods outperform TM methods. In fact, the result of combining the best method of each category does not bring better results than LM adaptation on its own. The best LM adaptation method is log-linear combination and using additional data from Open Subtitles is not effective neither in the TM nor in the LM.

The picture for the opposite direction is different: both LM and TM adaptation methods are useful, and the combination of them is clearly the best approach. The best translation model adaptation strategy is the combination of data selection (including OST) for the phrase table and fill-up for the reordering table,[12] which allows us to take advantage of the most relevant sentences from OST and avoid the overhead caused by its size.

The difference between both directions is probably related to the fact that Greek is a highly inflected language. When Greek is the SL, the out-of-domain parallel data helps to reduce the OOV rate and thus TM adaptation is necessary and data from OST is useful. When Greek is the TL, on the contrary, the LM is a very important part of the PBMT system: the number of translations of each English phrase is higher and a powerful LM helps to correctly combine the different hypotheses.

5 Neural machine translation

We also experimented with NMT in order to draw some preliminary conclusions about its performance in our target domain. We trained an NMT system for each direction from the in-domain data. We backtranslated (Sennrich et al., 2016a) the monolingual in-domain corpora with the best PBMT system for each direction in order to create additional parallel corpora and segmented the words by means of byte pair encoding (Sennrich et al., 2016b).[13] Our NMT systems follow the

[12] We did not apply data selection to the corpus from which the reordering table is built because fill-up has been reported to produce smaller models than linear interpolation (Bisazza et al., 2011).

[13] 60 000 join operations on the concatenation of SL and TL corpora.

Direction	BLEU	TER	chrF1
English-to-Greek	**0.2703**	**0.6362**	**0.6618**
Greek-to-English	0.3179	0.5826	**0.5554**

Table 5: Results of the NMT evaluation.

encoder-decoder architecture with attention proposed by Bahdanau et al. (2015)[14] and we followed the same training strategy as Sánchez-Cartagena and Toral (2016). Results, displayed in Table 5 show that, when systems are trained only on in-domain data, NMT outperforms PBMT by a statistically significant margin for the English-to-Greek direction (when the difference with the PBMT system trained only on in-domain data is statistically significant, a score is shown in bold).

6 Concluding remarks

We evaluated the use of web crawling, NMT and domain adaptation methods in PBMT in order to build English↔Greek MT systems for the tourism/culture domain. Data selection methods for the training data and a log-linear combination of LMs allowed us to build PBMT systems that perform better than those built from the simple concatenation of data, although the best set-up varies across directions. Preliminary experiments with NMT and in-domain data showed positive results, which encourage us to experiment with domain adaptation for NMT (Luong and Manning, 2015).

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (AbuMaTran).

References

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, United Kingdom.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

[14]We used the code available at: `https://github.com/sebastien-j/LV_groundhog/tree/master/experiments/nmt`

Bisazza, Arianna, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, USA.

Durrani, Nadir, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013. Edinburgh's machine translation systems for european language pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 114–121, Sofia, Bulgaria, August. Association for Computational Linguistics.

Heafield, Kenneth. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Koehn, Philipp and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 388–395, Barcelona, Spain.

Lewis, William and Sauleh Eetemadi. 2013. Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 281–291, Sofia, Bulgaria, August. Association for Computational Linguistics.

Luong, Minh-Thang and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 76–79, Da Nang, Vietnam, December.

Moore, Robert C. and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Papavassiliou, Vassilis, Prokopis Prokopidis, and Gregor Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth*

Workshop on Building and Using Comparable Corpora, pages 43–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Pecina, Pavel, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Aleš Tamchyna, Andy Way, and Josef van Genabith. 2015. Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, 49(1):147–193.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September.

Prokopidis, Prokopis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel global voices: a collection of multilingual corpora with citizen media stories. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Rubino, Raphael, Antonio Toral, Víctor M. Sánchez-Cartagena, Jorge Ferrández-Tordera, Sergio Ortiz Rojas, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Andy Way. 2014. Abu-matran at wmt 2014 translation task: Two-step data selection and rbmt-style synthetic rules. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 171–177, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Ruiz, Nicholas, Arianna Bisazza, Roldano Cattoni, and Marcello Federico. 2012. FBK's machine translation systems for IWSLT 2012's TED lectures. In *Proceedings of the 9th International Workshop on Spoken Language Translation*, pages 61–68, Hong Kong.

Sánchez-Cartagena, Víctor M. and Antonio Toral. 2016. Abu-matran at wmt 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation*, pages 362–370, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, USA.

Stolcke, Andreas et al. 2002. Srilm-an extensible language modeling toolkit. In *Interspeech*, volume 2002, page 2002.

Tiedemann, Jrg. 2012. Parallel data, tools and interfaces in opus. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Toral, Antonio, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way. 2016. Crawl and crowd to bring machine translation to under-resourced languages. *Language Resources and Evaluation*, pages 1–33.

Turner, Rochelle. 2015. Travel and Tourism Economic Impact 2015, Greece. `https://www.wttc.org/-/media/files/reports/economic%20impact%20research/countries%202015/greece2015.pdf`. Accessed: 2016-09-02.

Watanabe, Taro, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 764–773, Prague, Czech Republic.