The TALP–UPC Spanish–English WMT Biomedical Task: Bilingual Embeddings and Char-based Neural Language Model Rescoring in a Phrase-based System

Marta R. Costa-jussà, Cristina España-Bonet, Pranava Madhyastha, Carlos Escolano, José A. R. Fonollosa
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{marta.ruiz,jose.fonollosa}@upc.edu, {cristinae,pranava}@cs.upc.edu,
carlos.escolano@est.fib.upc.edu

Abstract

This paper describes the TALP–UPC system in the Spanish–English WMT 2016 biomedical shared task. Our system is a standard phrase-based system enhanced with vocabulary expansion using bilingual word embeddings and a character-based neural language model with rescoring. The former focuses on resolving out-of-vocabulary words, while the latter enhances the fluency of the system. The two modules progressively improve the final translation as measured by a combination of several lexical metrics.

1 Introduction

Machine Translation (MT) has been evolving in recent years achieving successful translations as shown by international evaluations such as WMT[1] and increasing use of MT in commercial applications. However, specific domains like legal, biomedical, etc., still lag behind the state-of-the-art MT systems. This can mostly be attributed to the lack of available corpora. The new biomedical task from WMT 2016 especially helps in improving our understanding in this direction.

In this paper, we describe our participation in the WMT 2016 biomedical task. We participated with a phrase-based SMT system enhanced with bilingual word embeddings and a character-based neural language model. Section 2 presents some related work to our approach. Next, Section 3 introduces the theoretical aspects of the system components and Section 4 the experiments. Finally, we justify our choice for the final submission and draw the conclusions in Section 5.

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

2 Related Work

In this paper, we are interested in research in the area that target OOVs and approaches to re-rank n-best lists of translations.

Our work closely follows Vulic and Moens (2015) and Zhao et al. (2015) in spirit, where word vectors are used to induce bilingual lexicons of words or phrases. We go a step further and build lexicons from bilingual word embeddings to be later used within an SMT system.

There is also a rich body of recent literature that focuses on obtaining bilingual word embeddings using aligned corpora (Bhattarai, 2012; Gouws et al., 2015; Kočiskỳ et al., 2014). We approach the problem differently and obtain embeddings separately on monolingual corpora and then use supervision in the form of a small sparse bilingual dictionary. This is similar to Mikolov et al. (2013b), who obtain monolingual embeddings for both the languages separately and then learn transformation for projecting the embeddings of words onto embeddings of the word translation pairs using a big bilingual dictionary.

On the other hand, there have been several language models used for rescoring in SMT. For example, neural feed-forward language models (Schwenk et al., 2006) have been used to rescore both n-gram-based and phrase-based systems. Mikolov (2012) re-ranks n-best lists with recurrent neural networks. Vaswani et al. (2013) combine feed-forward language models, with rectified linear units and noise-contrastive estimation. Luong et al. (2015) propose to use deeper neural models which improve re-ranking. In this paper, we are using Kim et al. (2016) a character-based language model to re-rank the output of the phrase-based system.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 463–468,
Berlin, Germany, August 11-12, 2016.

3 The Translation System

The TALP-UPC translation system is built on three different components. We describe their theoretical basis in the following subsections.

3.1 Phrase-based SMT

The standard phrase-based machine translation system (Koehn et al., 2003) focuses on finding the most probable target sentence given the source sentence. The phrase-based system has evolved from the noisy-channel to the log-linear model which combines a set of feature functions in the decoder, including the translation and language model, the reordering model and the lexical models. Although the phrase-based system is a commoditized technology used at the academic and commercial level, there are still many challenges to solve, such as OOVs.

3.2 Vocabulary Expansion using Bilingual Word-Embeddings

We look at this task as a bilinear prediction task as proposed by (Madhyastha et al., 2014). The proposed model makes use of word embeddings of both languages with no additional features. The basic function is formulated —the probability of a target word given a source word— as log-linear model and takes the following form:

$$\Pr(t|s;W) = \frac{\exp\{\ \tilde{s}(s)\ \ W\ \ \tilde{t}(t)\}}{_t\ \exp\{\ \tilde{s}(s)\ \ W\ \ \tilde{t}(t\ )\}} \quad (1)$$

Where $(.)$ denotes the n-dimensional distributed representation of the words, and we assume we have both source ($\tilde{s}$) embeddings and target ($\tilde{t}$) embeddings.

Essentially, our problem reduces to: a) first getting the corresponding word embeddings of the vocabularies on both the languages on a significantly large monolingual corpus and b) estimating W given a relatively small dictionary. To learn W we use the source word to target word dictionaries as training supervision.

We learn W by minimizing the negative log-likelihood of the dictionary using a nuclear norm regularized objective as: $L(W) = -\ _{s,t} \log(\Pr(t|s;W)) +\ \ W\ \ .\ \ $ is the constant that controls the capacity of W. To find the optimum, we follow the previous work and use an optimization scheme based on Forward-Backward Splitting (FOBOS) (Singer and Duchi, 2009).

Table 1: Size of the parallel (top) and monolingual (bottom) corpora used to train the translation systems

Corpus	Segments	Words	Vocab
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$
Bio-mono/en	$0.1 \cdot 10^6$	$2 \cdot 10^6$	$0.1 \cdot 10^6$
Bio-mono/es	$0.01 \cdot 10^6$	$0.1 \cdot 10^6$	$0.01 \cdot 10^6$
Wikipedia/en	$92 \cdot 10^6$	$1900 \cdot 10^6$	$2.0 \cdot 10^6$
Wikipedia/es	$20 \cdot 10^6$	$465 \cdot 10^6$	$0.8 \cdot 10^6$

3.3 Character-based Neural Language Model

Language models based on Recurrent Neural Networks are currently one of the best performing approaches in terms of perplexity (Mikolov et al., 2010). They are also a good re-ranking option in tasks such as speech recognition and machine translation. However, the standard lookup-based word embeddings are limited to a finite-size vocabulary for both computational and sparsity reasons. Moreover, the orthographic representation of the words is completely ignored. The standard learning process is blind to the presence of stems, prefixes, suffixes and any other kind of affixes in words.

As a solution to those drawbacks, new alternative character-based word embeddings have been recently proposed for tasks as language modeling (Kim et al., 2016; Ling et al., 2015), parsing (Ballesteros et al., 2015) or part-of-speech tagging (Ling et al., 2015; Santos and Zadrozny, 2014). For our system we selected the best character-based embedding architecture proposed by Kim et al. (Kim et al., 2016). The computation of the representation of each word starts with a character-based embedding layer that associates each word (sequence of characters) with a sequence of vectors. This sequence of vectors is then processed with a set of 1D convolution filters of different lengths (from 1 to 7 characters) followed with a max pooling layer and two additional highway layers. The output of the second highway layer provide us with the final vector representation of each source word that replaces the standard source word embedding in the recurrent neural network used for language modeling (Kim et al., 2016).

4 Experimental Framework

4.1 Data

Our main corpus is the compilation of the corpora assigned for the shared task, which was built using scientific publications gathered from the Scielo database. We focus on the Spanish–English language pair, for which the size of the corpora is summarised in Table 1. We further increase the vocabulary of the system by using standard parallel corpora for the Spanish–English language pair (i.e., UN corpora, Europarl corpora, News corpus, etc.[2]). This corpus appears as Quest in Table 1. For the monolingual corpus we use an English and Spanish Wikipedia dump[3].

The corpora has been pre-processed with a standard pipeline for both Spanish and English: tokenizing and keeping parallel sentences between 1 and 80 words. Additionally, for Spanish we used Freeling (Padró and Stanilovsky, 2012) to tokenize pronouns from verbs (i.e. comenzándose to comenzando + se), we also split prepositions and articles, i.e. del to de + el and al to a + el. This was done for similarity to English.

We divided the provided parallel corpus into training, development and test sets. Sentences from development and test set were taken randomly, proportionally to the amount of Medline and Scielo (biomedical and health) sources and only from unique parallel sentences.

Since the domain of the test set is the same as the domain of training corpus, the number of OOV words is small. Table 2 shows the total number and percentage of unknown words in our in-house development and test sets with respect to translation tables (see the following section). For comparison, we also include the figures for the two test sets made available for the final evaluation.

4.2 System Description

As introduced in the previous section, three different modules build our system: the SMT engine, the module to resolve OOVs and the module for re-reranking.

SMT Engine. Three different state-of-the-art phrase-based SMT translation systems are trained on the parallel corpora detailed in Table 1. For the purely in-domain system, we use only the biomedical data made available for the task (STT systems, small translation table). For more general systems, we also use the Quest data; we name these systems BTT (big translation table).

For the in-domain system, a 5-gram language model is estimated on the target side of the corpus using interpolated Kneser-Ney discounting with `SRILM` (Stolcke, 2002) (SLM, small language model). For the extended systems, we use all the monolingual corpora available and the target side of the large parallel corpus (BLM, big language model). Word alignment is done with `GIZA++` (Och and Ney, 2003) and both phrase extraction and decoding are done with the `Moses` package (Koehn et al., 2007). The optimisation of the weights of the model is trained with MERT (Och, 2003) against the BLEU (Papineni et al., 2002) evaluation metric on devBio.

OOVs resolution. This module first obtains bilingual embeddings from the monolingual ones as explained in Section 3.2. For estimating monolingual word vector models, we use the CBOW algorithm as implemented in the `Word2Vec` package (Mikolov et al., 2013a) using a 5-token window. We obtain 300 dimension vectors for English and Spanish from the monolingual and the source side of the parallel corpora in Table 1. The bilingual counterpart has been estimated using 34,806 words from the Apertium bilingual dictionary[4] as seed lexicon divided for training and validation. Each bilingual pair has an associated probability given by Eq. 1. We keep the top-10 pairs for each out-of-vocabulary word in the test (development) set and include these new translation options at decoding time. Since we are only dealing with OOVs, the new options do not interact with the other phrase pairs in the translation table, but there is interaction with the language model.

Re-ranking. The 1000-best list of translations given by the SMT engine is re-ranked using the characted-based language model described in Section 3.3. It has 1D convolutional filters of width [1,2,3,4,5,6,7] and size [50, 100, 150, 200, 200, 200, 200] for a total of 1,100 filters with a tanh activation, 2 highway layers with a ReLU activation, and 2 LSTM with 650 hidden units. The network

[2] In particular, we use the parallel data given for the Quality Estimation task at WMT13, `http://statmt.org/~buck/wmt13qe/wmt13qe_t13_t2_MT_corpus.tgz`

[3] Dumps downloaded from `https://dumps.wikimedia.org` in January 2015.

[4] `http://repositori.upf.edu/handle/10230/17110`

Table 2: Figures –tokens and OOVs– on the development and test sets used in the experiments

		English			Spanish		
	Seg.	Tokens	OOV_{STT}	OOV_{BTT}	Tokens	OOV_{STT}	OOV_{BTT}
devBio	1000	18967	16 (0.08%)	2 (0.01%)	19931	14 (0.07%)	6 (0.03%)
testBio	1000	26105	31 (0.11%)	19 (0.07%)	27651	25 (0.09%)	9 (0.03%)
Biological	4344	115709	434 (0.37%)	333 (0.29%)	126008	415 (0.33%)	254 (0.20%)
Health	5111	125624	133 (0.10%)	98 (0.08%)	146368	160 (0.11%)	40 (0.03%)

has been trained on the monolingual part of the in-domain data (Biomedical corpus in Table 1).

4.3 Results

We evaluate the performance of each module when added to the three standard SMT systems built with different amount of training data (STTSLM, STTBLM, BTTBLM). In the following, we denote the module for OOV resolution with **oov** and the module for re-reraking with **reranked**. For the **total.reranked** system, we re-ranked the n-best lists for the thirteen systems with our neural language model. We conduct the evaluation automatically with a set of lexical metrics calculated with the Asiya toolkit[5] (Giménez and Màrquez, 2010). Table 3 reports the results for the English-to-Spanish translation systems and Table 4 for the Spanish-to-English ones.

The first thing to notice is that the best translation is obtained when only in-domain data are used to build the translation model. This is true in both directions. When going from Spanish into English, we obtain 0.45 BLEU points of improvement when adding the oov module to the in-domain system (STTSLM.oov) and an additional 0.15 with the re-ranking module (STTSLM.oov.reranked). Even if the number of OOV is only a 0.09% in this test set, the improvement with this module is consistent through all metrics. The main reason is that making available new translation options at decoding time allows the language model to modify the sentence as a whole, and the neighbouring words can be modified accordingly.

In the English-to-Spanish direction, the trends are less homogeneous through the set of metrics. For BLEU and METEOR (with the stemming variant, MTRst), the best system is still STTSLM.oov. However, with NIST and TER, the best system is STTBLM. In this case, enlarging the language model has a similar effect as injecting new vocabulary through OOV translations. This is because only a 31% of the OOV belong to the biomedical domain, suggesting that in this case and for an in-domain test set, it is important to gain fluency on the general domain phrases. The effect of the re-ranking module is more evident in this direction: the more data one uses, the more distinct the final n-best list is and the more improvement one can obtain. For the in-domain system the re-ranking is not promoting a better translation, but for the general system the improvement is significant.

5 Conclusions

We have built thirteen translation systems per direction. The ones chosen for the final submission follow two criteria: i) they have a top performance according to BLEU and METEOR (the official metrics) and, ii) they allow us a coherent comparison among languages and methodologies. With this criteria, our primary submission both for the health and biological test sets is the strictly in-domain system with the OOV module (**STTSLM.oov**). For comparison, we also submitted our baseline as a second run: the same system without the OOV module (**STTSLM**). Finally, we submitted as third run a system with re-ranking of a 1000-best list. Due to time constraints, we could not submit the system that re-ranks all the n-best lists for the thirteen systems, **total.reranked**, but we used instead the two most promising options per direction.

According to the preliminary results of the shared task, the OOV module consistently improves the translations with respect to our baseline specially in the health subdomain as measured by BLEU. The effect is similar to the results in our in-house test set. On the other hand, the re-ranking module is also always better than the in-domain phrase-based baseline and, in this case, the performance on the competition test set is significantly better than the one in our test set, espe-

[5] http://nlp.cs.upc.edu/asiya

Table 3: Automatic evaluation of the in-house test set for the En2Es systems

	WER	PER	TER	BLEU	NIST	GTM-2	MTRst	MTRpa	RG-S*	ULC
BTTBLM.oov	48.45	29.82	44.27	43.84	8.81	36.30	61.58	62.87	49.85	66.03
BTTBLM.oov.reranked	47.58	**29.74**	43.56	44.43	8.90	36.97	62.01	63.25	50.43	67.16
BTTBLM	47.74	30.39	43.72	43.61	8.86	36.51	61.50	62.76	49.98	66.19
BTTBLM.reranked	47.64	29.91	43.52	44.24	8.89	36.90	61.88	63.14	50.29	66.95
STTBLM.oov	48.00	29.60	43.73	44.32	8.87	36.65	62.13	63.32	50.12	66.88
STTBLM.oov.reranked	47.22	29.85	43.11	44.57	8.96	37.21	62.22	63.42	50.44	67.57
STTBLM	**47.01**	29.93	**42.81**	44.51	**8.98**	37.36	62.28	63.47	50.49	67.75
STTBLM.reranked	47.10	29.91	42.96	44.65	8.97	37.40	62.31	63.46	**50.68**	67.78
STTSLM.oov	47.84	29.28	43.61	**44.99**	8.88	37.36	**62.33**	63.44	50.51	67.60
STTSLM.oov.reranked	47.41	29.82	43.25	44.52	8.94	37.29	62.25	63.36	**50.68**	67.54
STTSLM	47.29	29.84	43.16	44.64	8.96	**37.58**	62.27	63.42	50.56	67.71
STTSLM.reranked	47.40	29.93	43.24	44.39	8.94	37.36	62.21	63.30	50.56	67.44
total.reranked	47.06	29.82	43.03	44.75	**8.98**	37.56	**62.33**	**63.53**	50.66	**67.88**

Table 4: Automatic evaluation of the in-house test set for the Es2En systems

	WER	PER	TER	BLEU	NIST	GTM-2	MTRst	MTRpa	RG-S*	ULC
BTTBLM.oov	50.95	29.98	46.79	40.94	8.59	35.02	35.03	37.28	49.13	65.30
BTTBLM.oov.reranked	50.41	29.75	46.23	41.58	8.65	35.52	35.25	37.48	49.50	66.24
BTTBLM	50.21	29.33	45.98	41.97	8.68	35.88	35.44	37.65	50.01	66.97
BTTBLM.reranked	50.41	29.63	46.28	41.62	8.65	35.51	35.27	37.53	49.50	66.29
STTBLM.oov	50.75	29.95	46.68	40.82	8.61	34.83	35.05	37.12	49.15	65.27
STTBLM.oov.reranked	50.19	29.22	46.04	42.10	8.71	35.72	35.57	37.65	49.95	67.04
STTBLM	50.91	29.74	46.74	41.16	8.62	34.97	35.33	37.40	49.39	65.67
STTBLM.reranked	50.27	**29.08**	46.01	42.19	8.72	35.79	35.62	**37.66**	50.08	67.20
STTSLM.oov	**49.79**	29.45	**45.62**	42.16	**8.75**	**35.94**	35.57	37.60	**50.13**	**67.31**
STTSLM.oov.reranked	50.15	**29.08**	45.99	**42.30**	8.71	35.88	**35.65**	**37.66**	50.10	67.30
STTSLM	50.62	29.53	46.46	41.71	8.65	35.47	35.46	37.48	49.71	66.34
STTSLM.reranked	50.25	29.12	46.04	42.13	8.70	35.76	35.59	37.62	49.97	67.09
total.reranked	50.06	29.42	45.93	42.06	8.71	35.80	35.47	37.65	49.93	67.00

cially for English-to-Spanish. Run 3, the system that includes re-ranking with a char-based neural language model, is 2 points of BLEU over the average value among participants in the biological subdomain and 1 point of BLEU on the health subdomain.

Acknowledgements

This work is supported by the 7th Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951) and also by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, contract TEC2015-69266-P (MINECO/FEDER, UE).

References

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal, September.

Alexandre Klementiev Ivan Titov Binod Bhattarai. 2012. Inducing Crosslingual Distributed Representations of Words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India, December.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: an Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 748–756, July.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*.

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. arXiv preprint arXiv:1405.0947.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session, pages 177–180, June.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1520–1530, Lisbon, Portugal, September.

Thang Luong, Michael Kayser, and Christopher D. Manning. 2015. Deep neural language models for machine translation. In Proceedings of the Nineteenth Conference on Computational Natural Language Learning, pages 305–309, Beijing, China, July.

Swaroop Pranava Madhyastha, Xavier Carreras, and Ariadna Quattoni. 2014. Learning task-specific bilexical embeddings. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 161–171.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR. http://code.google.com/p/word2vec.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. abs/1309.4168.

Tomas Mikolov. 2012. Statistical Language Models based on Neural Networks. Ph.D. thesis, Brno University of Technology.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of the Association for Computational Linguistics, pages 160–167, Sapporo, Japan, July 6-7.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In International Conference on language Resources and Evaluation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the Association of Computational Linguistics, pages 311–318.

Cicero D. Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1818–1826.

Holger Schwenk, Marta R. Costa-Jussà, and José A. R. Fonollosa. 2006. Continuous space language models for the IWSLT 2006 task. In 2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006, pages 166–173.

Yoram Singer and John C Duchi. 2009. Efficient learning using forward-backward splitting. In Advances in Neural Information Processing Systems, pages 495–503.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of the Seventh International Conference of Spoken Language Processing (ICSLP2002), pages 901–904, Denver, Colorado, USA.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1387–1392, Seattle, Washington, USA, October.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, pages 719–725, July.

Kai Zhao, Hany Hassan, and Michael Auli. 2015. Learning Translation Models from Monolingual Continuous Representations. In The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015, pages 1527–1536, June.