Dublin City University at the TweetMT 2015 Shared Task

Dublin City University en la tarea TweetMT 2015

Antonio Toral, Xiaofeng Wu, Tommi Pirinen,
Zhengwei Qiu, Ergun Bicici, Jinhua Du
ADAPT Centre, School of Computing, Dublin City University, Ireland
{atoral, xwu, tpirinen, zhengwei.qiu2, ebicici, jdu}@computing.dcu.ie

Resumen: Describimos nuestra participación en TweetMT para tres pares de lenguas en ambas direcciones: castellano hacia/desde catalán, euskera y portugués. Hacemos uso de varias técnicas: traducción automática estadística y basada en reglas, segmentación de morfemas, selección de datos con ParFDA y combinación de sistemas. En cuanto a recursos, adquirimos grandes cantidades de tuits para llevar a cabo una adaptación de dominio monolingüe. Nuestro sistema ha sido el mejor de todos los enviados para cinco de los seis pares de lenguas.
Palabras clave: traducción automática, tuits, segmentación de morfemas, selección de datos

Abstract: We describe our participation in TweetMT for three language pairs in both directions: Spanish from/to Catalan, Basque and Portuguese. We used a range of techniques: statistical and rule-based MT, morph segmentation, data selection with ParFDA and system combination. As for resources, our focus was on crawling vast amounts of tweets to perform monolingual domain adaptation. Our system was the best of all systems submitted for five out of the six language directions.
Keywords: machine translation, tweets, morph segmentation, data selection

1 Introduction and Objectives

While statistical machine translation (SMT) can be considered a mature technology nowadays, one of its requirements is the availability of considerable amounts of parallel text for the language pair of interest. Ideally, the parallel text to train an SMT system should come from the same domain and genre as the text the system is going to be applied to. Thus, using MT to translate types of text for which no parallel data is available constitutes a challenge. This is the case for tweets and social media in general, the target text of the TweetMT shared task.

The main objective of our participation in the TweetMT 2015 shared task was to build the best MT systems for tweets we could with a clear constraint, i.e. it had to be done in a very short period and, to a large extent, be limited to available resources. We have taken part for three language pairs in both directions: Spanish (ES) from/to Catalan (CA), Basque (EU) and Portuguese (PT).

We decided to focus on making the best possible use of available techniques, tools and resources. Regarding techniques and tools, we rely on state-of-the-art SMT, morph segmentation for morphologically rich languages (EU), data selection with ParFDA for fast development of accurate SMT systems (Biçici, Liu, and Way, 2015) and domain adaptation (Biçici, 2015), the use of available open-source rule-based systems and, finally, system combination to take advantage of the strengths of the different systems we built. As for resources, we crawl vast amounts of tweets to perform monolingual domain adaptation and complement this with publicly available general-domain monolingual and parallel corpora.

The rest of the paper is organised as follows. Sections 2 and 3 detail the systems built and the resources used, respectively. Section 4 presents the evaluation and, finally, Section 5 outlines conclusions and lines of future work.

2 Architecture and Components of the System

Here we describe the components used in our translation pipeline. First, we pre-process the datasets (Section 2.1), then we use a set

of MT systems (Section 2.2) that can incorporate additional functionality (Sections 2.3 and 2.4). Finally, we combine MT systems (Section 2.5).

2.1 Data Preprocessing

Prior to be used, all the datasets used in our systems are preprocessed, as follows:

1. Punctuation normalisation, with Moses' (Koehn et al., 2007) script.
2. Sentence splitting and tokenisation, with Freeling (Padró and Stanilovsky, 2012).
3. Normalisation (only for tweets). We sort the vocabulary of a tweet corpus by word frequency and inspect the words that occur in at least 0.5% of the tweets, creating rules to convert informal words to their formal equivalent. This leads to just a handful of rules. E.g. in Spanish, "q", occurring in 2.62% of the tweets, is converted to its formal equivalent "que".
4. Truecasing, with a modified version of Moses' script. We added a set of start-of-sentence characters commonly used in Spanish: "-", "—", "¿", "“" and "‘".

2.2 MT Systems

We build SMT systems using two paradigms: phrase-based with Moses (Koehn et al., 2007) and hierarchical with cdec (Dyer et al., 2010). In both cases we use default settings. We also use off-the-shelf open-source rule-based MT (RBMT) systems. Namely, Apertium (Forcada et al., 2011) for ES↔CA, ES↔PT and EU→ES,[1] and Matxin (Mayor et al., 2011) for ES→EU.[2]

The SMT systems use 5-gram LMs with Knesser-Ney smoothing (Kneser and Ney, 1995) except for ParFDA Moses SMT systems, which use LMs of order 8 to 10. We build LMs on individual monolingual corpora (cf. Section 3.2) and interpolate them with SRILM (Stolcke and others, 2002) to minimise the perplexity on the dev set. Each target language and its corpora used to build LMs together with their interpolation weights are shown in Table 4. We observe that tweets are given very high weights even if they are not the biggest corpora in the mixes.

[1]Revisions 60356, 60384, and 60356, respectively.

[2]API at `http://ixa2.si.ehu.es/glabaka/Matxin.xml`

2.3 Morphological Segmentation

Morphological segmentation is a popular method to deal with SMT for morphologically differing languages by simply splitting words into sub-word units. The main benefits of morphological segmentation are to reduce the out-of-vocabulary (OOV) rate and to increase the percentage of 1 to 1 word alignments between morphosyntactically different languages; e.g. in our case, by matching inflectional suffixes in EU to syntactic prepositions in ES, we expect to improve the MT quality for the EU–ES language pair. The segmentation and de-segmentation is able to create word-forms not present in the training data by matching a translated stem with a correct suffix.

In our participation, morphological segmentation was only used for EU–ES on the EU side, since EU's morphology is significantly more complex than that of ES. For the remaining languages of the shared task, there is no such big difference in morphology complexity (all of them are closely-related as they belong to the same family) so the expected gains do not outweigh the added complexity of segmentation.

We use unsupervised statistical segmentation as provided by Morfessor 2.0 Baseline (Virpioja et al., 2013).[3] The basic setup for segmentation is the same as in the AbuMaTran project submission to the WMT 2015 translation task (Rubino et al., 2015). However, some minor Twitter-related preprocessing has been added in order to keep URLs and hashtags intact. The parameters used for Morfessor training are the default of version 2.0.2-alpha and the data for training is the EU side of the ES–EU parallel training data (cf. Section 3.1).

To gauge the effects of our method as well as the morphological complexity of EU as compared to ES we show in Table 1 the OOV rates and vocabulary sizes of the ES and EU sides of the ES–EU training corpus, and EU corpora after morphological segmentation. Segmentation reduces the type-to-token ratio by a factor of 6 and the OOV rate by almost a factor of 10.

2.4 ParFDA

ParFDA parallelizes instance selection with an optimized parallel implementation of

[3]`http://www.cis.hut.fi/projects/morpho/morfessor2.shtml`

Corpora	Tokens	Types	OOV
ES	30,532,489	296,612	14.5 %
EU	24,966,862	605,207	25.4 %
EU morphs	35,293,220	100,990	2.6 %

Table 1: Size of ES–EU training corpus in word tokens (ES and EU sides) and in morph tokens (EU).

5-gram	OOV				perplexity			
$S \rightarrow T$	C train	FDA train	FDA LM	%red	C train	FDA train	FDA LM	%red
CA–ES	2948	2957	2324	.21	332	336	294	.11
EU–ES	3021	3046	2443	.19	462	483	546	-.18
PT–ES	2871	2896	1951	**.32**	633	623	486	.23
ES–CA	3338	3345	2890	.13	325	330	338	-.04
ES–EU	4110	4129	3349	.19	745	761	637[a]	.15[a]
ES–PT	3087	3117	2216	.28	993	941	746	**.25**

Table 2: LM comparison built from training corpus (C train), ParFDA selected training data (FDA train), ParFDA selected LM data (FDA LM). %red is reduction proportion.

[a]ES–EU LM is recomputed after the task, removing duplicates, which slightly decrease BLEU, increase NIST.

FDA5 and significantly reduces the time to deploy accurate SMT systems especially in the presence of large training data and still achieve state-of-the-art SMT performance (Biçici, Liu, and Way, 2015; Biçici and Yuret, 2015). Detailed composition of the available corpora, which is referred to as constrained (C), are provided in Section 3. For ES, we also included LDC Gigaword corpora (Ângelo Mendonça et al., 2011). The size of the LM corpora includes both the LDC and the monolingual LM corpora provided. ParFDA selected training and LM data obtains accurate translation outputs with the selected LM data reducing the number of OOV tokens by up to 32% and the perplexity by up to 25% and allows us to model higher order dependencies (Table 2).

2.5 System Combination

For each language direction we have built up to five systems, as detailed in Sections 2.2 to 2.4: (i) phrase-based and (ii) hierarchical SMT, (iii) phrase-based with morph segmentation, (iv) phrase-based with ParFDA and (v) RBMT. We hypothesise these systems to have complementary strengths, and thus we decide to perform system combination. To that end we use MEMT (Heafield and Lavie, 2010), with default settings, except for the parameter `length`, for which we use its default (7) for all directions except for ES→EU, for which we use 5 according to empirical results on the development set.

3 Resources Employed

3.1 Parallel Corpora

Ideally, we would use data in the same domain and genre as the test set, i.e. tweets. We have access to parallel tweets provided by the task for ES–CA and ES–EU (4,000 parallel tweets for each language pair, we use 1,000 for dev and the remaining 3,000 for training). For ES–PT we have access to 999 parallel tweets (we use them for dev) from Brazilator,[4] a recent project by DCU and Microsoft to translate tweets from the 2014 soccer World Cup across 24 language directions.

As the availability of parallel tweets for the language pairs of TweetMT 2015 is rather limited (at most we have 4,000 per language pair), we use additional sources of parallel data. For ES–CA we use elPeriodico (eP)[5] and a selection of contemporary novels. For ES–EU, translation memories (TMs) provided by the shared task[6] and two corpora from Opus (Tiedemann, 2012):[7] Open subtitles 2013 and Tatoeba. Finally, for ES–PT we use Europarl v7[8] and two corpora from Opus: news-commentary and Tatoeba. Table 3 provides details on these corpora.

3.2 Monolingual Corpora

Our main source of monolingual data is in-domain and comes from crawled tweets. We use TweetCat (Ljubešić, Fišer, and Erjavec, 2014) and crawl tweets for all the target languages (CA, ES, EU and PT) during March and April 2015.

For each language we create two lists of words as required by the crawler: (i) most common discriminating words (up to 100), these are words that are unique to the language and they are used to seed the crawler so that it can find candidate tweets; and (ii) most common words of the language (200), these are used to determine the language of

[4]`http://www.cngl.ie/brazilator`

[5]`http://catalog.elra.info/product_info.php?products_id=1122`

[6]`http://komunitatea.elhuyar.org/tweetmt/resources/`

[7]`http://opus.lingfil.uu.se/`

[8]`http://www.statmt.org/europarl/`

Pair	Corpus	# s.	# tokens
ES–CA	tweets	3K	48k, 48k
	eP	0.6M	13.5M, 14M
	novels	47K	.78M, .86M
ES–EU	tweets	3K	42K, 38K
	TMs	1.1M	28.9M, 23.5M
	OpenSubs	0.16M	1.2M, 1.0M
	Tatoeba	902	6.7K, 5.5K
ES–PT	EU	1.9M	54M, 53M
	NC	9K	.26M, .25M
	Tatoeba	53K	.42M, .41M

Table 3: Parallel corpora used for training. For each corpus we provide its number of sentence pairs (# s.) and tokens on both sides (# tokens).

Lang	Corpus	# tokens	Weights
CA	tweets	29M	0.60
	caWaC	0.5G	0.33
	eP	14M	0.07
ES	tweets	129.2M	0.75
	news	0.4G	0.21
	europarl	60M	0.04
EU	tweets	11.3M	0.97
	Wikipedia	11.5M	0.01
	TMs	23M	0.02
PT	tweets	33M	0.93
	Wikipedia	166M	0.02
	Others	286M	0.05

Table 4: Monolingual corpora used for training. For each corpus we show its number of tokens (# tokens) and its weight in LM interpolation.

crawled tweets. These two lists are derived from a list of the most common words found in a corpus of subtitles.[9]

The tweets crawled are post-processed with langid[10] to identify their language. We keep the tweets whose langid's confidence score is above a certain threshold, which is set empirically at 0.7 by inspecting tweets.

In addition to crawled tweets, we use the target sides of the parallel corpora (cf. Section 3.1 and a set of monolingual corpora as follows. For CA we use caWaC (Ljubešić and Toral, 2014), a corpus crawled from the `.cat` top level domain. For ES, news crawl and news-commentary from WMT'13.[11] For EU, a dump from Wikipedia (20150407). For PT, the news sources CETEMPublico,[12] and CETENFolha,[13] and a dump from Wikipedia (20150510).

Table 4 shows details on these corpora including their interpolation weights (cf. Section 2.2).

4 Evaluation

We report our results on the development set (all systems built) and then on the test set (systems submitted).

4.1 Evaluation on Development Data

Table 5 presents the results obtained on the devset by the individual systems and a set of combinations for the three language pairs we covered: ES–CA, ES–EU and ES–PT. The scores were obtained on raw MT output (i.e. tokenised and truecased) as calculated by us with BLEU (Papineni et al., 2002) (multibleu cased as included in Moses version 3) and TER (Snover et al., 2006) (as implemented in TERp version 0.1). Due to time constraints not all the possible combinations were tried. The scores of the best individual system and combination are shown in bold.

At least one of the combinations obtains better scores (both in terms of BLEU and TER) than the best individual system (except for ES↔PT with BLEU and for CA→ES with TER), supporting our hypothesis that the individual systems built are complementary. Although SMT systems outperform RBMT systems for all directions,[14] the addition of RBMT in system combinations has a positive impact (except for ES↔PT). Phrase-based SMT outperforms hierarchical SMT for related language pairs (ES–CA and ES–PT), but the opposite is true for the unrelated language pair ES–EU. We hypothesise this is due to the fact that ES and EU follow different word orders (SVO and SOV, respectively), and this leads to pervasive long reorderings in translation, that are better modelled with a hierarchical approach.

[9] `https://onedrive.live.com/?cid=3732e80b128d016f&id=3732E80B128D016F!3584`

[10] `https://github.com/saffsd/langid.py`

[11] `http://www.statmt.org/wmt13/`

[12] `http://www.linguateca.pt/cetempublico/`

[13] `http://www.linguateca.pt/cetenfolha/`

[14] When interpreting the results, it should be taken into account that automatic metrics are known to be biased towards statistical MT approaches (Callison-Burch, Osborne, and Koehn, 2006).

	System	BLEU	TER
ES→CA	Moses (1)	82.21	0.1102
	cdec (2)	81.45	0.1128
	ParFDA (3)	**82.37**	**0.1062**
	Apertium (4)	78.17	0.1310
	1+2	81.71	0.1102
	1+4	**82.37**	**0.1057**
	1+2+4	81.93	0.1085
CA→ES	Moses (1)	**82.52**	0.1086
	cdec (2)	81.76	0.1118
	ParFDA (3)	82.16	**0.1063**
	Apertium (4)	77.96	0.1329
	1+2	82.38	0.1088
	1+4	**82.58**	0.1077
	1+2+4	82.38	0.1083
	1+3+4	82.45	**0.1074**
ES→EU	Moses (1)	22.57	0.6116
	cdec (2)	**23.7**	**0.5863**
	ParFDA (3)	21.59	0.6181
	Matxin (4)	12.66	0.7436
	Morph (5)	5.20	0.8812
	1+2	23.18	0.5796
	1+4	18.36	0.6112
	1+2+4	23.58	0.5771
	1+2+4+5	24.07	**0.5741**
	1+2+3+4+5	**24.42**	0.5777
EU→ES	Moses (1)	24.21	0.6228
	cdec (2)	**24.65**	**0.5911**
	ParFDA (3)	22.25	0.6346
	Apertium (4)	18.36	0.6918
	Morph (5)	11.25	0.9655
	1+2	24.18	0.5883
	1+4	24.33	0.6076
	1+2+4	24.94	0.5831
	1+2+4+5	**25.21**	**0.5792**
ES→PT	Moses (1)	**29.21**	0.6052
	cdec (2)	28.14	**0.5962**
	ParFDA (3)	27.74	0.6164
	Apertium (4)	24.96	0.6272
	1+2	**28.76**	0.5891
	1+4	26.58	0.6082
	1+2+4	27.00	**0.5878**
PT→ES	Moses (1)	**30.47**	0.5267
	cdec (2)	29.42	**0.5254**
	ParFDA (3)	29.63	0.5338
	Apertium (4)	27.52	0.5335
	1+2	**29.9**	0.5230
	1+4	30.01	0.5131
	1+2+4	29.89	**0.5089**

Table 5: Results on the dev set.

	System	BLEU	TER
ES→CA	DCU1 (1+4)	0.7669	0.1740
	DCU2 (1)	**0.7899**†	**0.1626**†
	DCU3 (1+2+4)	0.7630	0.1738
CA→ES	DCU1 (1+4)	0.7826	0.1506
	DCU2 (1+2+4)	0.7816	0.1500
	DCU3 (1+3+4)	**0.7943**†	**0.1431**†
ES→EU	DCU1 (1+2+4)	0.2455	0.6533
	DCU2 (1+2+3+4+5)	**0.2636**†	**0.6469**†
	DCU3 (1+2+4+5)	0.2493	0.6553
EU→ES	DCU1 (2)	0.2687	0.6512
	DCU2 (1+2+4)	0.2698	0.6406
	DCU3 (1+2+4+5)	**0.2728**	**0.6363**
ES→PT	DCU1 (1)	0.3595	0.5290
	DCU2 (1+2)	**0.3711**†	**0.5157**†
	DCU3 (1+2+4)	0.3687	0.5163
PT→ES	DCU1 (1)	0.4465	0.5767
	DCU2 (1+2)	0.4467	0.5627
	DCU3 (1+2+4)	**0.4524**†	**0.5403**†

Table 6: Results on the test set.

4.2 Evaluation on Test Data

Table 6 presents the results on the test set of the systems we submitted. The scores shown are the ones reported by the organisers (case-insensitive BLEU and TER) on post-processed MT outputs (detokenised and detruecased). For each language direction we submitted the three systems that obtained the best performance on the dev set. The scores of the best submitted system are shown in bold.

Out of six directions, our best submission is the top performing system for five of them (indicated with †). For most directions, the addition of a RBMT system leads to better performance. Similarly, for the directions where we have used segmentation (ES↔EU) and ParFDA (CA→ES and ES→EU), the addition of systems based on these techniques had a positive impact on the results.

We now delve deeper into the results obtained by SMT systems based on ParFDA (cf. Section 2.4). Although ParFDA systems were submitted to the shared task only as part of system combinations, we have evaluated *a posteriori* the performance of this technique by means of standalone systems on the test set. ParFDA Moses SMT system obtains top results in CA→ES and ES→CA and close to top results in other language pairs with 1.21 BLEU points average difference to the top (Table 7). An interesting feature of

TweetMT	CA–ES	EU–ES	PT–ES
ParFDA	**.8012**	.2713	.4374
Top	.7942	.3109	.4519
diff	-.007	.0396	.0145
LM order	8	8	8
	ES–CA	ES–EU	ES–PT
ParFDA	**.7926**	.2482	.3589
Top	.7907	.2636	.3711
diff	-.0019	.0154	.0122
LM order	8	10	8

Table 7: BLEU results for ParFDA standalone systems on the test set, their difference to the top, and ParFDA LM order used. ParFDA obtains top results in CA→ES and ES→CA and 1.21 BLEU points average difference.

ParFDA regards its ability to build and deploy SMT systems in a quick manner. In the specific case of TweetMT, ParFDA took about 8 hours to build for ES→CA and 28 hours for PT→ES taking about 11 GB and 27 GB disk space in total, respectively.

5 Conclusions and Future Work

This paper has described our participation in the TweetMT 2015 shared task. Our focus has been on rapid development of MT systems adapted to tweets by making the best possible use of available techniques, tools and resources. Our best submissions have been the ones that combine different MT systems (except for ES→CA), supporting our hypothesis that the techniques we have used are complementary.

As for future work, we consider several possible avenues. First, we would like to analyse in detail the translations produced by our systems in order to derive findings beyond the ones we can extract from the automatic evaluation metrics used in the task. Second, most of the tweets in the test set use formal language,[15] and thus we would like to test our systems in a more representative set of tweets where informal language would be expected to be more pervasive.

Acknowledgments

This research is supported by the EU 7th Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran), by SFI as part of the ADAPT research center (07/CE/I1142) at Dublin City University and the project "Monolingual and Bilingual Text Quality Judgments with Translation Performance Prediction" (13/TIDA/I2740). We also thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support. Finally, we would like to thank Mikel L. Forcada and Iacer Calixto for their advice on normalising tweets for Basque and Portuguese, respectively, and Gorka Labaka for his help with Matxin's API.

[15]This is due to the fact that they are extracted from twitter accounts that publish tweets in multiple languages, and such accounts belong, to a large extent, to institutions that use formal language.

References

Ângelo Mendonça, Daniel Jaquette, David Graff, and Denise DiPersio. 2011. Spanish Gigaword third edition, Linguistic Data Consortium.

Biçici, Ergun. 2015. Domain adaptation for machine translation with instance selection. *The Prague Bulletin of Mathematical Linguistics*, 103:5–20.

Biçici, Ergun, Qun Liu, and Andy Way. 2015. ParFDA for fast deployment of accurate statistical machine translation systems, benchmarks, and statistics. In *Proceedings of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Biçici, Ergun and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP)*, 23:339–350.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

Dyer, Chris, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the*

Association for Computational Linguistics (ACL).

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Heafield, Kenneth and Alon Lavie. 2010. Combining machine translation output with open source: The carnegie mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93:27–36.

Kneser, Reinhard and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec. 2014. TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Ljubešić, Nikola and Antonio Toral. 2014. cawac - a web corpus of catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz de Ilarraza Sánchez, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. *Matxin*, an open-source rule-based machine translation system for basque. *Machine Translation*, 25(1):53–82.

Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Rubino, Raphael, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz-Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. 2015. Abu-MaTran at WMT 2015 Translation Task: Morphological Segmentation and Web Crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

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 Association for machine translation in the Americas*, pages 223–231.

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

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), 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).

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.