Deep evaluation of hybrid architectures: simple metrics correlated with human judgments

Gorka Labaka, Arantza Díaz de Ilarraza,
Kepa Sarasola
University of the Basque Country
gorka.labaka@ehu.es,
jipdisaa@ehu.es, kepa.sarasola@ehu.es

Cristina España-Bonet,
Lluís Màrquez
Universitat Politècnica de Catalunya
cristinae@lsi.upc.edu,
lluism@lsi.upc.edu

Abstract

The process of developing hybrid MT systems is guided by the evaluation method used to compare different combinations of basic subsystems. This work presents a deep evaluation experiment of a hybrid architecture that tries to get the best of both worlds, rule-based and statistical. In a first evaluation human assessments were used to compare just the single statistical system and the hybrid one, the rule-based system was not compared by hand because the results of automatic evaluation showed a clear disadvantage. But a second and wider evaluation experiment surprisingly showed that according to human evaluation the best system was the rule-based, the one that achieved the worst results using automatic evaluation. An examination of sentences with controversial results suggested that linguistic well-formedness in the output should be considered in evaluation. After experimenting with 6 possible metrics we conclude that a simple arithmetic mean of BLEU and BLEU calculated on parts of speech of words is clearly a more human conformant metric than lexical metrics alone.

1 Introduction

The process of developing hybrid MT systems is guided by the evaluation method used to compare different combinations of basic subsystems. Direct human evaluation is more accurate but unfortunately it is extremely expensive, so automatic metrics have to be used in prototype developing. However the method should evaluate different systems with the same criteria, and these criteria should be as close as possible to human judgment.

It is well known that rule-based and phrase-based statistical machine translation paradigms (RBMT and SMT, respectively) have complementary strengths and weaknesses. First, RBMT systems tend to produce syntactically better translations and deal with long distance dependencies, agreement and constituent reordering in a better way, since they perform the analysis, transfer and generation steps based on syntactic principles. On the bad side, they usually have problems with lexical selection due to a poor handling of word ambiguity. Also, in cases in which the input sentence has an unexpected syntactic structure, the parser may fail and the quality of the translation decrease dramatically. On the other side, phrase-based SMT models usually do a better job with lexical selection and general fluency, since they model lexical choice with distributional criteria and explicit probabilistic language models. However, phrase-based SMT systems usually generate structurally worse translations, since they model translation more locally and have problems with long distance reordering. They also tend to produce very obvious errors, which are annoying for regular users, e.g., lack of gender and number agreement, bad punctuation, etc. Moreover, SMT systems can experience a severe degradation of performance when applied to corpora different from those used for training (*out-of-domain* evaluation).

It is also well known that the BLEU metric (Papineni et al., 2002) is actually the most used metric in statistical MT. But several doubts have arisen around BLEU (Melamed et al., 2003; Callison-Burch et al., 2006; Koehn and Monz, 2006). In addition to the fact that it is extremely difficult to interpret what is being expressed in BLEU (Melamed et al., 2003), improving its value neither guarantees an improvement in the translation quality (Callison-Burch et al., 2006) nor offers as much correlation with human judgment as was believed (Koehn and Monz, 2006). Those problems have also been detected when translating to Basque (Mayor, 2007; Labaka, 2010).

In the last few years, several new evaluation metrics have been suggested to consider a higher level of linguistic information (Liu and Gildea, 2005; Popović and Ney, 2007; Chan and Ng, 2008), and different methods of metric combination have been tested. Due to its simplicity, we decided to use the idea presented by Giménez and Màrquez (2008), where the different simple metrics are combined by means of the arithmetic mean.

In this work we present some surprising results we have achieved in a deep evaluation of a hybrid architecture. In a first step we used human evaluation to compare just the single statistical system and the hybrid one, we did not compare the rule-based system by hand because the results of automatic evaluation showed a clear disadvantage. But a second and wider evaluation experiment surprisingly showed that according to human evaluation the best system was the rule-based, the one that achieved the worst results using automatic evaluation. We tried to make a diagnosis of this phenomenon, and then based on this we finally found a simple but more human conformant metric that we plan to use in training new versions of our hybrid system.

In the next section of this paper we describe the hybrid system. Section 3 presents the evaluation experiments: the corpora used in them, the first experiment comparing just the single statistical system and the hybrid one, and the second and wider evaluation experiment which compares the all three systems. Then Section 4 describes the process of searching for other automatic metrics being more human conformant. And finally, the last section is devoted to conclusions and future work.

2 The hybrid system, SMatxinT

Statistical Matxin Translator, SMatxinT in short, is a hybrid system controlled by the RBMT translator and enriched with a wide variety of SMT translation options (España-Bonet et al., 2011).

The two individual systems are a rule-based Spanish-Basque system called Matxin (Alegria et al., 2007) and a standard phrase-based statistical MT system based on Moses which works at the morpheme level allowing to deal with the rich morphology of Basque (Labaka, 2010).

The initial analysis of the source sentence is done by Matxin. It produces a dependency parse tree, where the boundaries of each phrase are marked. In order to add hybrid functionality two new modules are introduced to the RBMT architecture (Figure 1): the tree enrichment module, which incorporates SMT additional translations to each phrase of the syntactic tree; and a monotonous decoding module, which is responsible for generating the final translation by selecting among RBMT and SMT partial translation candidates from the enriched tree.

The tree enrichment module introduces two types of translations for the syntactic constituents given by Matxin: 1) the SMT translation(s) of every phrase, and 2) the SMT translation(s) of the entire subtree containing that phrase. For example, the analysis of the test fragment "*afirmó el consejero de interior*" (said the Secretary of interior) gives two phrases: the head "*afirmó*" (said) and its children "*el consejero de interior*" (the Secretary of interior). The full rule-based translation is "*Barne Sailburua baieztatu zuen*" and the full SMT translation is "*esan zuen herrizaingo sailburuak*". SMatxinT considers these two phrases for the translation of the full sentence, but also the SMT translations of their constituents ("*esan zuen*" and "*herrizaingo sailburuak*"). However, short phrases may have a wrong SMT translation because of a lack of context. To overcome this problem SMatxinT also uses the translation of a phrase extracted from a longer SMT translation ("*herrizaingo sailburuak*" in the previous example). So, in order to translate "*afirmó el consejero de interior*" the system has produced 5 distinct phrases, a number that can be increased by considering a n-best list of SMT outputs.

After tree enrichment, the transfer and generation steps of the RBMT system are carried out in a usual way, and a final monotonous decoder chooses among the options. A key aspect for the performance of the system is the election of the features

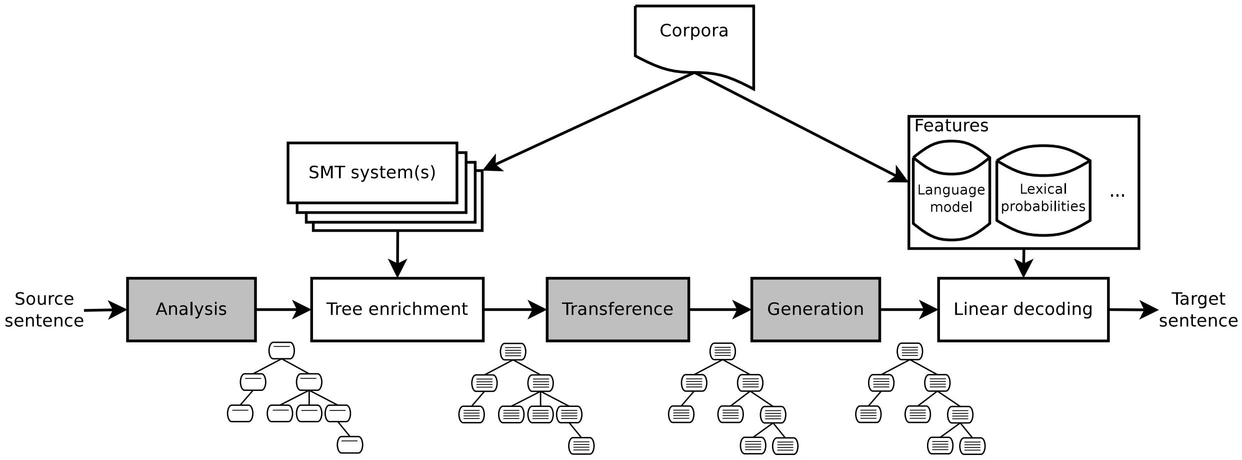

Figure 1: General architecture of SMatxinT. The RBMT modules which guide the MT process are the grey boxes.

for this decoding. The results we present here are obtained with a set of eleven features. Three of them are the usual SMT features (language model, word penalty and phrase penalty). We also include four features to show the origin of the phrase and the consensus among systems (a counter indicating how many different systems generated the phrase, two binary features indicating whether the phrase comes from the SMT/RBMT system or not, and the number of source words covered by the phrase generated by both individual systems simultaneously). Finally, we use the lexical probabilities in both directions in two forms: a similar approach to IBM-1 probabilities modified to take unknown alignments into account and a lexical probability inferred from the RBMT dictionary. We refer the reader to España-Bonet et al. (2011) for further details.

3 Experiments

In our experiments we evaluate both individual systems and the final hybrid: SMT, Matxin and SMatxinT. The language pair of application is dictated by the rule-based system and, in this case, Matxin works with the Spanish-to-Basque translation. Basque and Spanish are two languages with very different morphologies and syntaxes.

3.1 Bilingual and monolingual corpora

The corpus built to train the SMT system consists of four subsets: (1) six reference books translated manually by the translation service of the University of the Basque Country (EHUBooks); (2) a collection of 1,036 articles published in Spanish and Basque by the Consumer Eroski magazine[1] (Consumer); (3) translation memories mostly using administrative language developed by Elhuyar[2] (ElhuyarTM); and (4) a translation memory including short descriptions of TV programmes (EuskaltelTB). Table 1 shows some statistics on the corpora, giving some figures about the number of sentences and tokens.

		sentences	tokens
EHUBooks	Spanish Basque	39,583	1,036,605 794,284
Consumer	Spanish Basque	61,104	1,347,831 1,060,695
ElhuyarTM	Spanish Basque	186,003	3,160,494 2,291,388
EuskaltelTB	Spanish Basque	222,070	3,078,079 2,405,287
Total	Spanish Basque	491,853	7,966,419 6,062,911

Table 1: Statistics on the bilingual collection of parallel corpora.

The training corpus is then basically made up of administrative documents and descriptions of TV programs. For development and testing we extracted some administrative data for the *in-domain* evaluation and selected one collection of news for the *out-of-domain* study, totaling three sets:

Elhuyardevel and *Elhuyartest*: 1,500 segments each, extracted from the administrative documents.

[1] http://revista.consumer.es
[2] http://www.elhuyar.org/

NEWStest: 1,000 sentences collected from Spanish newspapers with two references.

Additionally, we collected a 21 million word monolingual corpus, which together with the Basque side of the parallel bilingual corpora, builds up a 28 million word corpus. This monolingual corpus is also heterogeneous, and includes text from two sources of news: the Basque corpus of Science and Technology (ZT corpus) and articles published by Berria newspaper (Berria corpus).

3.2 First experiment of evaluation

According to the automatic evaluation, carried out in the previous article and extended in Table 4, the rule-based Matxin system is clearly the worst system obtaining the worst scores for both metrics (BLEU and TER) in both test corpora. On the other hand, the evaluation of the hybrid system varies depending on the test set. On the in-domain corpora (Elhuyar test set), the BLEU score achieved by SMatxinT is slightly worse than the scores obtained by the single SMT system, but better according to TER (Snover et al., 2006) evaluation. The distinct behavior between metrics and the small differences do not allow us to define a clear preference between statistical and hybrid systems. On the contrary, on the out-domain corpora (NEWS test set), SMatxinT consistently archives better scores than any other system.

Based on these results, we stated that the low in-domain performance of the Matxin penalizes the hybrid system, preventing it to overcome the single SMT system. But, in the out-domain test set, where the scores of Matxin were not so far from the rest of the systems, our hybridization technique was able to combine the best of both systems obtaining the best translation. In order to verify this assertion, we carried out an human evaluation, where we asked four evaluators to determine the preference between the hybrid and the SMT translations of 100 sentences randomly chosen from the NEWS test set. The figures obtained corroborated that the hybrid system outperforms the single SMT system in the out-domain corpora.

3.3 Deeper evaluation: Human evaluation to compare the three systems

In order to get a more detailed insight of the performance of our systems, we recently extended this manual evaluation to the rest of the systems and test corpora. That way, we selected another 100 sentences from the Elhuyar test set and asked the same four evaluators to assess the preference between the three system pairs (SMT-Matxin, SMT-SMatxinT, Matxin-SMatxinT).

Surprisingly, according to this manual evaluation the best system is the rule-based Matxin system, the worst ranked one using automatic evaluation. Even for in-domain evaluation it is clearly better than the statistical system and of similar quality as the hybrid one, that is slightly superior to the statistical system. For out-domain evaluation the differences are very clear: the rule-based Matxin system clearly outperforms the hybrid system and this one outperforms the statistical system.

This can be seen in Table 2. The table shows the number of times that a system is better than the other for those sentences where there was full agreement among evaluators (*Agreement*) and for the full subset (*All*). Results are given for the three system pairs on the two test sets, the in-domain and the out-of-domain ones.

We confirmed these surprising results of manual evaluation by examining some examples where BLEU scores did not reflect the difference of quality between translation outputs. Let us analyze the example shown in Table 3, that is, the translation of the source sentence *"Legasa cuenta ya con un convenio sobre la recuperación de bienes comunales."*. The table shows the source sentence with its meaning in English together with two translation references and the output given by the two individual systems.

In this example, the output of the rule-based system is adequate, but BLEU is unable to recognize some linguistic equivalences: *jadanik* and *jada* are synonymous, as well as *berreskuratzearen inguruan* and *berreskuratze gainean.* Similarly *herri ondasunak* and *herri-ondasunen* are almost the same because the "-" is optional, and using *Legasa* instead of *Legasak* is a common error easy to understand. The following segments are quasi equivalents: *hitzarmena du* and *kontatzen du hitzarmen batekin*. All these correspondences are trivial for humans but invisible for the BLEU metric.

On the other hand, the output of the statistical system is harder to understand. By using *Legasako* instead of *Legasak*, the sentence becomes difficult to

			System 1	**Tied**	**System 2**
Elhuyar (in-domain)	SMT vs. SMatxinT	Agreement	5 (9.4%)	**31 (58.5%)**	17 (32.1%)
		All	25 (12.5%)	**109 (54.5%)**	66 (33.0%)
	SMT vs. **Matxin**	Agreement	14 (23.7%)	19 (32.2%)	**26 (44.1%)**
		All	41 (20.5%)	79 (39.5%)	**80 (40.0%)**
	SMatxinT vs. Matxin	Agreement	19 (28.8%)	**24 (36.4%)**	23 (34.8%)
		All	59 (29.5%)	**82 (41.0%)**	59 (29.5%)
NEWS (out-domain)	SMT vs. **SMatxinT**	Agreement	15 (21.4%)	22 (31.4%)	**33 (47.2%)**
		All	40 (20.0%)	74 (37.0%)	**86 (43.0%)**
	SMT vs. **Matxin**	Agreement	11 (17.7%)	13 (21.0%)	**38 (61.3%)**
		All	32 (16.0%)	64 (32.0%)	**104 (52.0%)**
	SMatxinT vs. **Matxin**	Agreement	19 (26.4%)	13 (18.1%)	**40 (55.5%)**
		All	49 (24.5%)	54 (27.0%)	**97 (48.5%)**

Table 2: Manual evaluation for random subset of 100 sentences of each test corpus.

Source	*Legasa cuenta ya con un convenio sobre la recuperación de bienes comunales.*
(English)	*Legasa already has a convention on the recovery of community property.*
Ref. 1	*Legasak hitzarmena du jada herri ondasunak berreskuratzearen inguruan.*
Ref. 2	*Legasak badauka ondasun komunalak berreskuratzeari buruzko hitzarmena.*
Matxin	*Legasa jadanik kontatzen du hitzarmen batekin herri-ondasunen berreskuratze gainean.*
SMT	*legasako hitzarmena du dagoeneko berreskuratzeari buruzko ondasunak komunalak.*
SMatxinT	*dagoeneko legasako hitzarmena berreskuratzeari buruzko ondasun komunalak*

Table 3: Example where an understandable translation obtained by Matxin is penalized by BLEU, but the confusing SMT translation gets a good BLEU score.

understand, and the same happens with the strange end of the sentence. However, this translation obtains a good evaluation score because every word but one is in the references.

4 Searching for human conformant automatic metrics

In view of the large difference between the results obtained by standard automatic metrics and the manual evaluation, and considering that the human evaluators value syntactical correctness more than the common lexical metrics (such as BLEU and TER) do, we considered the possibility of using metrics that use a higher level of linguistic information (Liu and Gildea, 2005; Popović and Ney, 2007; Giménez and Màrquez, 2007; Chan and Ng, 2008). Thus, in addition to the standard BLEU and TER, we applied these same metrics over the sequences of syntactic categories, parts of speech (PoS), resulting BLEU_PoS and TER_PoS. Table 4 shows how the metrics that use linguistic information obtain more similar results to those achieved by the manual evaluation. Thus, in our out-domain evaluation the metrics that use PoS information show the same preference between systems than the human assessment. That is, Matxin gets the best results, followed by SMatxinT and SMT. Similarly, in the in-domain test set, the human preference of SMatxinT over the statistical system is clearer with this type of metrics. Despite this, PoS based metrics can not fully compensate the high penalty that Matxin receives and this system remains the lowest ranked in the Elhuyar test set (in-domain), although the distance is shorter.

However, those results are provably biased by the fact that both SMT and SMatxinT systems are optimized to rise their BLEU score. Thus, they get a high lexical matching to the reference, at the expense of the syntactical correctness. Similarly, the use of metrics that only take into account a even more specific aspect of translation, such as the coincidence of PoS, are not suitable to be used as the unique metric for the whole developing cycle. Using such metrics on SMT parameter optimization, for example, could lead to get translations whose lexical correction is fully ignored. So this kind of met-

		BLEU	**TER**	**BLEU_PoS**	**TER_PoS**	**comb_BLEU**	**comb_all**
Elhuyar (in-domain)	**Matxin**	5.25	84.51	25.63	52.82	15.44	7.88
	SMT	**14.53**	71.60	30.78	48.82	22.65	11.53
	SMatxinT	14.48	**70.50**	**31.96**	**47.07**	**23.22**	**11.82**
Elhuyar (in-domain) hand evaluated sentences	**Matxin**	5.85	84.95	26.68	52.19	16.27	8.29
	SMT	12.75	75.58	30.15	49.37	21.45	**11.38**
	SMatxinT	**13.37**	**75.09**	**31.39**	**48.63**	**22.38**	**11.38**
NEWS (out-domain)	**Matxin**	11.65	72.39	**39.19**	**42.40**	**25.42**	**12.93**
	SMT	14.45	70.18	31.09	48.65	22.77	11.59
	SMatxinT	**15.08**	**67.72**	34.55	45.56	24.82	12.62
NEWS (out-domain) hand evaluated sentences	**Matxin**	11.01	73.55	**38.74**	**43.07**	**24.88**	**12.65**
	SMT	11.32	73.08	29.56	50.49	20.44	10.41
	SMatxinT	**13.64**	**70.42**	35.34	46.82	24.49	12.45

Table 4: Automatic scores of all individual and hybrid systems.

rics should be combined with metrics that also take into account other aspects of the translation, as lexical matching. In the literature different methods of metric combination have been tested. Among other methods, one can find those based on linear combinations (Padó et al., 2009; Liu and Gildea, 2007; Giménez and Màrquez, 2008), regression based algorithms (Paul et al., 2007; Albrecht and Hwa, 2008) or a variety of supervised machine learning algorithms (Quirk et al., 2005; Amigó et al., 2005).

Due to its simplicity and the results achieved, we decided to use the idea presented by Giménez and Màrquez (2008), where the different metrics are combined just by means of the arithmetic mean. This method of combination, despite its simplicity, obtained competitive results on the MetricsMATR shared task (Callison-Burch et al., 2010). Thus we have defined two metrics that combine lexical information with PoS information: (1) one that combines the four metrics (BLEU, TER, BLEU_PoS and TER_PoS) we tested and (2) another one that combines only BLEU with BLEU_PoS.

BLEU and BLEU_PoS are quality measures (higher score means higher quality) while TER and TER_PoS are error measure (lower score means higher quality). Due to the different nature of the metrics and to be able to combine all of these four metrics by means of the arithmetic mean, we had to modify the values of TER to become quality measures. Thus, the new metrics are calculated using the following formulas:

$$Comb\_BLEU = (BLEU + BLEU\_PoS)/2$$

$$\begin{aligned} Comb\_All = & (BLEU + BLEU\_PoS + (100 - TER) \\ & + (100 - TER\_PoS))/4 \end{aligned}$$

The two metrics that combine lexical metrics with PoS information obtained results similar to those based only on PoS, in terms of preference between systems. In the same way, BLEU_PoS and TER_PoS, Comb_BLEU and Comb_All established the same preference order as the manual evaluation, except in the case of Matxin in the in-domain test set. But, unlike those metrics based only on PoS information, the combined metrics are more suitable as they allow a better syntactic adequacy while they maintain correct lexical matchings.

In addition to this correlation at the document level, we also wanted to check the correlation of each metric at sentence level where manual assessments were set. For each sentence in which both human assessments agree, we have compared the result with the preference for each metric. To define which is the preference for each metric, we considered that the automatic metric prefers a translation if one of the translations gets a score 10% higher than the other. In cases where the relative difference is not higher than 10%, we consider that the automatic metric is not able to discriminate between the two translations. Table 5 shows the percentage of sentences where each automatic metric's preference coincides with the one set by both human evaluators (we discard the cases in which human evaluations have not agreed).

		BLEU	**TER**	**BLEU_PoS**	**TER_PoS**	**comb_BLEU**	**comb_all**
Elhuyar (in-domain)	SMT vs. SMatxinT	34 (64%)	33 (62%)	33 (62%)	30 (56%)	**35 (66%)**	31 (58%)
	SMT vs. Matxin	23 (39%)	23 (39%)	25 (42%)	22 (37%)	24 (41%)	**26 (44%)**
	SMatxinT vs. Matxin	25 (38%)	**29 (44%)**	**29 (44%)**	27 (41%)	25 (38%)	28 (42%)
NEWS (out-domain)	SMT vs. SMatxinT	35 (50%)	31 (44%)	36 (51%)	**38 (54%)**	**38 (54%)**	34 (49%)
	SMT vs. Matxin	31 (50%)	29 (47%)	**42 (68%)**	38 (61%)	39 (63%)	**42 (68%)**
	SMatxinT vs. Matxin	38 (53%)	38 (53%)	39 (54%)	36 (50%)	**46 (64%)**	36 (50%)

Table 5: Sentence by sentence correlation between human evaluation and automatic metrics.

These figures show that the metrics based on linguistic information (both, those that only uses PoS information and those that combine it with lexical information) get more coincidences than those that only use lexical information (BLEU or TER).

5 Conclusions

In this work we present an in-depth evaluation of SMatxinT, a hybrid system that is controlled by the RBMT translator and enriched with a wide variety of SMT translation options. The results of the human evaluation, where the translation of the two individual systems and SMatxinT were compared in pairs, established that Matxin, the RBMT system, achieved the best performance followed by SMatxinT, while the SMT system generated the worst translations.

Those results, very far from what the automatic metrics (BLEU and TER) show, corroborate the already known inadequacy of the metrics that measure only the lexical matching for comparing systems that use so different translation paradigms. This kind of metrics are biased in favor of the SMT, as it happens in our evaluation, where the statistical system achieves the best results in the in-domain evaluation, even when it generates the worst translations according to the manual assessment.

To address these limitations of the metrics that are only based on lexical matching, we defined a couple of metrics that seek to ensure the syntactic correctness, calculating the same expressions but at the PoS level. These metrics, which are able to assess the syntactic correctness, have shown a higher level of agreement with human assessments both at document and sentence level.

Nevertheless, the metrics that assess specific aspects of the translation (such as PoS matching) do not ensure the absolute quality of the translation, and should be combined with regular lexical matching metrics. At the time of combining these metrics, we opted for simplicity and we used the arithmetic mean. This method, despite its simplicity, has already shown its suitability before.

Our combined metrics are simple and able to maintain a higher correlation with manual evaluation than the usual lexical metrics, while ensure the lexical matching.

We are planning to use this simple combination of metrics in developing new versions of our hybrid system. Simultaneously we are adapting linguistic tools to the Asiya Open Toolkit[3] to test other new evaluation metrics that consider a higher level of linguistic information.

Acknowledgments

This work has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247914 (MOLTO project, FP7-ICT-2009-4-247914), the Spanish Ministry of Science and Innovation (OpenMT-2 project, TIN2009-14675-C03-01) and the Local Government of the Basque Country (Berbatek project, IE09-262).

References

Joshua Albrecht and Rebecca Hwa. 2008. Regression for machine translation evaluation at the sentence level. *Machine Translation*, pages 1–27.

Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeri Mayor, and Kepa Sarasola. 2007. Transfer-based MT from spanish into basque: Reusability, standardization and open source. *Lecture Notes in Computer Science*, 4394:374–384.

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 2005. Qarla: a framework for the evalua-

[3] http://nlp.lsi.upc.edu/asiya

tion of text summarization systems. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 280–289, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics (EACL)*, pages 249–256.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio, June. Association for Computational Linguistics.

Cristina España-Bonet, Gorka Labaka, Arantza Díaz de Ilarraza, Lluis Màrquez, and Kepa Sarasola. 2011. Hybrid machine translation guided by a rulebased system. In *Proceedings MT Summit XIII*, Xiamen, China, Septemper.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 256–264, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *In Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, pages 102–121.

Gorka Labaka. 2010. *EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation*. Ph.D. thesis, University of the Basque Country.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 25–32.

Ding Liu and Daniel Gildea. 2007. Source-language features and maximum correlation training for machine translation evaluation. In *HLT-NAACL'07*, pages 41–48.

Aingeru Mayor. 2007. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. Ph.D. thesis, Euskal Herriko Unibertsitatea.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 61–63, Morristown, NJ, USA. Association for Computational Linguistics.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 297–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, 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 of the Association for Computational Linguistics*, pages 311–318.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2007. Reducing Human Assessments of Machine Translation Quality to Binary Classifiers. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Maja Popović and Hermann Ney. 2007. Word error rates: decomposition over pos classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 48–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231.