Chapter 6

Deep evaluation of hybrid architectures: Use of different metrics in MERT weight optimization

Cristina España-Bonet, Gorka Labaka, Arantza Díaz de Ilarraza, Lluís Màrquez, Kepa Sarasola

UPC Barcelona, University of the Basque Country

The process of developing hybrid MT systems is usually guided by an evaluation method used to compare different combinations of basic subsystems. This work presents a deep evaluation experiment of a hybrid architecture, which combines rule-based and statistical translation approaches. Differences between the results obtained from automatic and human evaluations corroborate the inappropriateness of pure lexical automatic evaluation metrics to compare the outputs of systems that use very different translation approaches. An examination of sentences with controversial results suggested that linguistic well-formedness should be considered in the evaluation of output translations. Following this idea, we have experimented with a new simple automatic evaluation metric, which combines lexical and PoS information. This measure showed higher agreement with human assessments than BLEU in a previous study (Labaka et al., 2011). In this paper we have extended its usage throughout the system development cycle, focusing on its ability to improve parameter optimization.

Results are not totally conclusive. Manual evaluation reflects a slight improvement, compared to BLEU, when using the proposed measure in system optimization. However, the improvement is too small to draw any clear conclusion. We believe that we should first focus on integrating more linguistically representative features in the developing of the hybrid system, and then go deeper into the development of automatic evaluation metrics.

6.1 Introduction

The process of developing hybrid MT systems is guided by the evaluation method used to compare outputs of 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 the outputs of 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).

Because of these complementary virtues and drawbacks several works are being devoted to build hybrid systems with components of both approaches. A classification and a summary of hybrid architectures can be seen in Thurmair (2009). The case we present here is within the philosophy of those systems where the RBMT system leads the translation and the SMT system provides complementary information. Following this line, Habash et al. (2009) enrich the dictionary of a RBMT system with phrases from an SMT system. Federmann et al. (2010) use the translations obtained with a RBMT system and substitute selected noun phrases by their SMT counterparts. Globally, their results improve the individual systems when the hybrid system is applied to translate into languages with a richer morphology than the source.

Regarding the evaluation of the final system and its components, still nowadays, the BLEU metric (Papineni et al., 2002) is the most used metric in MT, but several doubts have arisen around it (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 such high correlation with human judgment as was believed (Koehn and Monz, 2006).

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 a set of simple metrics are combined by means of the arithmetic mean.

This work presents a deep evaluation experiment of a hybrid architecture that tries to get the best of both worlds, rule-based and statistical. The results obtained corroborated the known doubts about BLEU. And suggests that the further development of the hybrid system should be guided by a linguistically more informed metric that should be able to capture the syntactic correctness of the rule-based translation, which is preferred by human assessors.

In the next section of this paper we describe the hybrid system. Section 6.3 presents the evaluation experiments: the corpora used in them, and the results of the automatic and manual evaluations. Finally, the last section is devoted to conclusions and future work.

6.2 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).

6.2.1 Individual systems

The two individual systems combined in SMatxinT are a rule-based Spanish–Basque system called Matxin (Mayor et al., 2011) 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).

Matxin is an open-source RBMT engine, whose main goal is to translate from Spanish into Basque using the traditional transfer model. Matxin consists of three main components: (i) analysis of the source language into a dependency tree structure; (ii) transfer from the source language dependency tree to a target language dependency structure; and (iii) generation of the output translation from the target dependency structure.

The engine reuses several open tools and it is based on an unique XML format for the flow between the different modules, which makes easier the interaction among different developers of tools and resources. The result is an open source software which can be downloaded from matxin.sourceforge.net, and it has an on-line demo[1] available since 2006.

For the statistical system, words are split into several morphemes by using a Basque morphological analyzer/lemmatizer, aiming at reducing the sparseness produced by the agglutinative nature of Basque and the small amount of parallel corpora. Adapting the baseline system to work at the morpheme level mainly consists of training the decoder on the segmented text. The SMT system trained on segmented words generates a sequence of morphemes. So, in order to obtain the final Basque text from the segmented output, a word-generation post-process is applied.

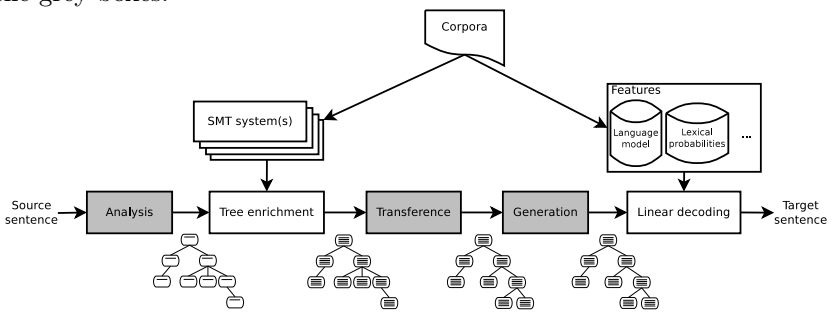
State-of-the-art tools are used in this case. GIZA++ toolkit (Och, 2003) is used for the alignments, SRILM toolkit (Stolcke, 2002) for the language model and the Moses Decoder (Koehn et al., 2007). We used a log-linear functions: phrase translation probabilities (in both directions), word-based translation probabilities (lexicon model, in both directions), a phrase length penalty and the target language model. The language model is a simple 3-gram language model with modified Kneser-Ney smoothing. We also used a lexical reordering

[1] `http://www.opentrad.com`

model ('msd-bidirectional-fe' training option). Parameter optimization was done following the usual practice, i.e., Minimum-Error-Rate Training (Och, 2003), however, the metric used for the optimization is not only BLEU, but it depends on the system as it will be seen.

6.2.2 Hybridisation

Figure 6.1: General architecture of SMatxinT. The RBMT modules which guide the MT process are the grey boxes.



The initial analysis of the source sentence is done by Matxin. It produces a dependency parse tree, where the boundaries of each syntactic phrase are marked. In order to add hybrid functionality two new modules are introduced to the RBMT architecture (Figure 6.1): 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 text 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. So, 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 the n-best lists.

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 for this decoding. The results we present here are obtained with a set of eleven features. Three of them are usually used as standard 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.

6.3 Experiments

The language pair used at evaluation 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 morphology and syntax.

In previous experiments we evaluated all systems by means of both automatic an manual evaluations (Labaka et al., 2011). Those results corroborated 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 achieved the best results in the in-domain evaluation, even when it generated 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 metric that seeks to check the syntactic correctness, calculating the same expressions but at the PoS level and combining it with lexical BLEU through the arithmetic mean. This metric, which is able to assess the syntactic correctness, has shown a higher level of agreement with human assessments both at document and sentence level.

But evaluation metrics are not only used for comparing different systems, those metrics are also used to guide the development of the systems. Thus, being aware of the problems of BLEU to identify many of the good translations generated by the RBMT system, we used linguistically informed metrics not only on the evaluation, but also in MERT optimization of the linear decoder. So, in addition to individual systems, we will evaluate three different hybrid systems, depending on the metric used in optimization (BLEU, METEOR and BLEU_c, a new defined metric according to Eq. 6.1).

6.3.1 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[2] (Consumer); (3) translation memories mostly using administrative language developed by Elhuyar[3] (ElhuyarTM); and (4) a translation memory including short descriptions of TV programmes (EuskaltelTB). All together they made up a corpus of 8 mil-

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

lion words in Spanish and 6 mil-ion words in Basque. Table 6.1 shows some statistics on the corpora, giving some figures about the number of sentences and tokens.

Table 6.1: Statistics on the bilingual collection of parallel corpora.

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

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 we selected a collection of news for the *out-of-domain* study, totaling three sets:

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

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: the Basque Corpus of Science and Technology (ZT corpus[4]) and articles published by Berria newspaper (Berria corpus).

6.3.2 Automatic Evaluation

In order to perform the automatic evaluation of the translations we use a subset of lexical metrics available in the Asiya evaluation package (Giménez and Màrquez, 2010). Tables 6.2 and 6.3 show the BLEU, TER and METEOR scores for the in-domain test set (Elhuyartest) and the out-of-domain one (NEWStest) respectively[5]. Besides, the tables include the score given by the combination of metrics for the two individual systems (Matxin and SMT) and three hybrid systems SMatxinT that have been optimized against these different metrics. Results of Google Translate[6] are given as control system.

In Labaka et al. (2011) it was shown that a simple combination of n-gram matching metrics at different linguistic levels, such as words and PoS, is more correlated with human

[4] www.ztcorpusa.net/

[5] Figures do not exactly match the ones presented in previous work, since we correct some capitalization errors.

[6] http://translate.google.com/

assessments than just the lexical match. Therefore, we use this new metric, BLEU_c, not only to evaluate the translations but also to optimize the system.

$$\text{BLEU}_c = (\text{BLEU} + \text{BLEU}_{PoS})/2 \tag{6.1}$$

Table 6.2: Automatic evaluation of the in-domain test set, Elhuyartest, for the individual and hybrid systems.

		BLEU	**METEOR**	**TER**	**BLEU$_c$**
Ind. systems	**Matxin**	6.07	27.20	83.49	19.65
	SMT	**16.50**	37.49	70.39	27.64
Control	**Google**	8.19	28.02	78.43	20.73
SMatxinT	**BLEU**	16.09	**38.24**	69.92	**27.95**
	BLEU$_c$	15.36	**38.24**	70.78	27.33
	METEOR	15.87	37.77	**67.77**	27.53

Table 6.3: Automatic evaluation of the out-of-domain test set, NEWStest, for the individual and hybrid systems.

		BLEU	**METEOR**	**TER**	**BLEU$_c$**
Ind. systems	**Matxin**	12.67	36.10	69.16	31.98
	SMT	15.84	37.70	66.52	31.01
Control	**Google**	12.36	32.57	70.44	29.08
SMatxinT	**BLEU**	16.61	39.24	64.50	32.77
	BLEU$_c$	**17.11**	**39.94**	63.84	**33.39**
	METEOR	16.76	39.30	**62.83**	32.50

According to all the automatic metrics Matxin is the worst system both for in-domain and out-of-domain data. The statistical system is worse than the hybrid models for out-of-domain data and shows a similar performance in the in-domain test set. In this case, the BLEU score achieved by SMatxinT is slightly worse than the scores obtained by the single SMT system, but better according to the rest of metrics. 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 (NEWStest), SMatxinT consistently achieves better scores than any other system.

The use of different metrics in the MERT optimization does not significantly affect the final evaluation. The systems that have been optimized with respect to different metrics obtained very similar results and, when these differences exists, they are not consistent between different evaluation test set or metrics.

In the in-domain evaluation, although the differences are small, the hybrid system optimized on BLEU gets the best results according to BLEU, METEOR and BLEUc. In contrast, the TER metrics assigns the best score to the hybrid system that is optimized on METEOR. It is worth noting that the optimizations on BLEUc and METEOR does not improve results by those metrics.

In the out-domain corpus, although the differences remain small, the results are more stable. In this test set, the hybrid system that achieves the best evaluation is the one optimized on BLEUc, improving the results obtained by the BLEU optimization according to all evaluation metrics. In this corpus, as in the in-domain one, the system optimized on

METEOR achieves results particularly high in the TER metric, which makes if to be the best system according to this metric.

Based on these results, one could state that the low in-domain performance of 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.

6.3.3 Human Evaluation

As in previous works, we contrast those automatic results with a manual evaluation carried out on 100 sentences randomly chosen from the in-domain test set (Elhuyartest) and another 100 sentences chosen from the out-domain test set (NEWStest). The human evaluators are asked to order the 5 translation provided (both individual systems and three different optimizations of SMatxinT). Human evaluators are allowed to determine that various translations are equally good. Depending on how many draws there are, the ranking scope can vary for 1 to 5 (when there is not any draw) to 1 to 1 (when all systems are considered equal). So, we normalized all rankings to the 0-1 scope (where 0 is the best system and 1 is the worst in all cases).

Table 6.4 shows the original and normalized average rankings obtained by each system. According those results, in the in-domain test set Matxin obtains the best ranking, but differences to the three SMatxinT instances are not significant. Those systems that use linguistically motivated metrics (METEOR and BLEU$_c$) in MERT obtain slightly better results than the instance optimized over BLEU. The SMT system, in turn, obtains the worst ranking. On the other hand, in the out-domain evaluation the differences are bigger: Matxin, the rule-based system, clearly outperforms the hybrid systems and these ones outperform the statistical system. The differences between different optimizations of SMatxinT are not significant.

Table 6.4: Real and normalized mean of the ranking manually assigned to each system.

		Elhuyartest		**NEWStest**	
		ranking	**norm.**	**ranking**	**norm.**
Ind. systems	**Matxin**	2.070	0.396	1.705	0.275
	SMT	2.510	0.532	2.605	0.625
SMatxinT	**BLEU**	2.165	0.423	2.210	0.485
	BLEU$_c$	2.085	0.399	2.110	0.445
	METEOR	2.095	0.403	2.125	0.470

Each sentence, 100 in each test set, has been assessed by two evaluators. Agreement between evaluators is difficult to check, as qualitatively small changes between them can produce multiple single changes in the precedence numbers in the ranking. For example, between the following two rankings

Matxin 1, BLEU 2, BLEU$_c$ 2, METEOR 2, SMT 3
Matxin 1, BLEU 2, BLEU$_c$ 3, METEOR 3, SMT 4

three precedence numbers are changed, but there is only a single qualitative difference (in

the second ranking the system trained with BLEU is better than those trained with BLEUc and METEOR).

To make the rankings more comparable we discretized the assigned ranking into 4 possible values: *best*, *intermediate*, *worst* and *all-draw*. The *best* and *worst* values mean that the system has been asserted as the best or the worst system. The *intermediate* value is assigned to other systems. In the cases that all systems are assigned to the same rank the *all-draw* value is assigned.

Table 6.5 shows the times that both evaluators assigned the same discrete ranking. Between brackets, the times that each evaluator assigns this ranking is shown. In some cases, the agreement is high, as when Matxin is claimed as the best out-domain system, *47(51+64)*. But generally the agreement is not very high.

Table 6.5: Discrete ranking results. Figures correspond to agreement of both evaluators, between brackets each evaluator's figures.

	Elhuyartest			
	best	intermediate	worst	all-draw
Matxin	24 (34+42)	9 (26+19)	20 (38+32)	0 (2+7)
SMT	9 (22+23)	7 (31+23)	30 (45+47)	0 (2+7)
BLEU	8 (27+19)	22 (52+43)	8 (19+31)	0 (2+7)
BLEU$_c$	12 (27+18)	29 (55+45)	7 (16+30)	0 (2+7)
METEOR	6 (28+19)	24 (54+47)	6 (16+27)	0 (2+7)
	NEWStest			
	best	intermediate	worst	all-draw
Matxin	47 (51+64)	4 (22+12)	10 (25+19)	0 (2+5)
SMT	7 (20+11)	6 (21+25)	41 (57+59)	0 (2+5)
BLEU	11 (28+15)	27 (44+43)	21 (26+37)	0 (2+5)
BLEU$_c$	12 (27+17)	28 (50+44)	15 (21+34)	0 (2+5)
METEOR	11 (26+16)	26 (46+42)	18 (26+37)	0 (2+5)

These results further demonstrate the equality of the systems, thickened by the lack of agreement between evaluators. In addition, it also shows some interesting results, as the fact that even in-domain the RBMT system produces more sentences tagged as the best translation. But the system also generates a high number of sentences labeled as the worst translation. So, in the overall assessment it fails to distance itself from the hybrid systems (which produce less 'best' translations, but also less 'worst' translations).

6.4 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 all the individual systems was ranked, 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 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 is 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 metric that seeks to ensure the syntactic correctness, combining lexical BLEU with PoS matching information. At the time of combining these metrics, we opted for simplicity and we used the arithmetic mean of BLEU in words and PoS. This method, despite its simplicity, has already shown its suitability before. Our combined metric is simple and able to maintain a higher correlation with manual evaluation than the usual lexical metrics, while ensures the lexical matching.

But evaluation metrics are not only used for comparing different systems, those metrics are also used to guide the optimization of the systems. In practical terms, in our hybrid architecture, we used those metrics to identify the features that are able to differentiate the best translation proposed by different approaches. Thus, being aware of the problems of BLEU to identify many of the good translations generated by the RBMT system, we used linguistically informed metrics not only on the evaluation, but also in MERT optimization of the linear decoder. So, in addition to individual systems, we evaluate three different hybrid systems, depending on the metric used in optimization. According to the results achieved, the use of different metrics in optimization has low impact in translation quality. Although the use of BLEU_c in optimization slightly improves the results achieved by manual evaluation, this improvement is too small to draw clear conclusions.

We consider that the minimal differences that exist between different optimizations are due to the lack of linguistic features at monotonous decoding. Current 11 features are mainly devoted to characterize the origin system of a given phrase and the probabilities for the lexical translation. In MERT optimization, the evaluation metrics are only used to find out which of the features present in the decoding are the most useful at generating the final translation. So, if there are no features which depend on the PoS in our case, or on higher level information such as the type of chunk, they may not be informative enough to strengthen the metric. In this case, optimization has little room for improvement.

Given these results, the need to provide more in-depth linguistic information to the evaluation metrics is undeniable. But, since we carry out our research in translation into Basque, we have at our disposal few linguistic tools, much less than for languages like English. Future work should first focus on integrating more representative linguistic features in the hybrid system which allow a qualitative leap in the translations quality. Then the small improvements reported here could be confirmed or ruled out.

Acknowledgments

This research has been partially funded by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03) and the EC Seventh Framework Programme under grant agreement numbers 247914 (MOLTO project, FP7-ICT-2009-4-247914) and 247762 (FAUST project, FP7-ICT-2009-4-247762).

Bibliography

Callison-Burch, Chris, 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.

Chan, Yee Seng 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: ACL.

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

Federmann, C., A. Eisele, Y. Chen, S. Hunsicker, J. Xu, and H. Uszkoreit. 2010. Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 77–81. Uppsala, Sweden: ACL.

Giménez, Jesús 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: ACL.

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

Habash, Nizar, Bonnie Dorr, and Christof Monz. 2009. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation* 23:23–63.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Prague, Czech Republic.

Koehn, Philipp and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the 1st Workshop on Statistical Machine Translation*, pages 102–121.

Labaka, Gorka. 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.

Labaka, Gorka, Arantza Díaz de Ilarraza, Cristina España-Bonet, Lluis Màrquez, and Kepa Sarasola. 2011. Deep evaluation of hybrid architectures: simple metrics correlated with human judgements. In *Proceedings of International Workshop on Using Linguistic Information for Hybrid Machine Translation LIHMT*. Barcelona.

Liu, Ding 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.

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

Melamed, I. Dan, 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: ACL.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167.

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

Popović, Maja 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: ACL.

Stolcke, A. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference of Spoken Language Processing*, vol. 2, pages 901–904.

Thurmair, G. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceddings of MT Summit XII*.