The Universitat d'Alacant hybrid machine translation system for WMT 2011

Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz
Transducens Research Group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071, Alacant, Spain
{vmsanchez,fsanchez,japerez}@dlsi.ua.es

Abstract

This paper describes the machine translation (MT) system developed by the Transducens Research Group, from Universitat d'Alacant, Spain, for the WMT 2011 shared translation task. We submitted a hybrid system for the Spanish–English language pair consisting of a phrase-based statistical MT system whose phrase table was enriched with bilingual phrase pairs matching transfer rules and dictionary entries from the Apertium shallow-transfer rule-based MT platform. Our hybrid system outperforms, in terms of BLEU, GTM and METEOR, a standard phrase-based statistical MT system trained on the same corpus, and received the second best BLEU score in the automatic evaluation.

1 Introduction

This paper describes the system submitted by the Transducens Research Group (Universitat d'Alacant, Spain) to the shared translation task of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation (WMT 2011). We participated in the Spanish–English task with a hybrid system that combines, in a phrase-based statistical machine translation (PBSMT) system, bilingual information obtained from parallel corpora in the usual way (Koehn, 2010, ch. 5), and bilingual information from the Spanish–English language pair in the Apertium (Forcada et al., 2011) rule-based machine translation (RMBT) platform.

A wide range of hybrid approaches (Thurmair, 2009) may be taken in order to build a machine translation system which takes advantage of a parallel corpus and explicit linguistic information from RBMT. In particular, our hybridisation approach directly enriches the phrase table of a PBSMT system with phrase pairs generated from the explicit linguistic resources from an Apertium-based shallow-transfer RBMT system. Apertium, which is described in detail below, does not perform a complete syntactic analysis of the input sentences, but rather works with simpler linear intermediate representations.

The rest of the paper is organised as follows. Next section overviews the two MT systems we combine in our submission. Section 3 outlines related hybrid approaches, whereas our approach is described in Section 4. Sections 5 and 6 describe, respectively, the resources we used to build our submission and the results achieved for the Spanish–English language pair. The paper ends with some concluding remarks.

2 Translation approaches

We briefly describe the rationale behind the PBSMT (section 2.1) and the shallow-transfer RBMT (section 2.2) systems we have used in our hybridisation approach.

2.1 Phrase-based statistical machine translation

Phrase-based statistical machine translation systems (Koehn et al., 2003) translate sentences by maximising the translation probability as defined by the log-linear combination of a number of feature functions, whose weights are chosen to opti-

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 457–463,
Edinburgh, Scotland, UK, July 30–31, 2011.

mise translation quality (Och, 2003). A core component of every PBSMT system is the phrase table, which contains bilingual phrase pairs extracted from a bilingual corpus after word alignment (Och and Ney, 2003). The set of translations from which the most probable one is chosen is built by segmenting the source-language (SL) sentence in all possible ways and then combining the translation of the different source segments according to the phrase table. Common feature functions are: source-to-target and target-to-source phrase translation probabilities, source-to-target and target-to-source lexical weightings (calculated by using a probabilistic bilingual dictionary), reordering costs, number of words in the output (word penalty), number of phrase pairs used (phrase penalty), and likelihood of the output as given by a target-language (TL) model.

2.2 Shallow-transfer rule-based machine translation

The RBMT process (Hutchins and Somers, 1992) can be split into three different steps: i) analysis of the SL text to build a SL intermediate representation, ii) transfer from that SL intermediate representation to a TL intermediate representation, and iii) generation of the final translation from the TL intermediate representation.

Shallow-transfer RBMT systems use relatively simple intermediate representations, which are based on lexical forms consisting of lemma, part of speech and morphological inflection information of the words in the input sentence, and apply simple shallow-transfer rules that operate on sequences of lexical forms: this kind of systems do not perform a full parsing. Apertium (Forcada et al., 2011), the shallow-transfer RBMT platform we have used, splits the transfer step into structural and lexical transfer. The lexical transfer is done by using a bilingual dictionary which, for each SL lexical form, always provides the same TL lexical form; thus, no lexical selection is performed. Multi-word expressions (such as *on the other hand*, which acts as a single adverb) may be analysed by Apertium to (or generated from) a single lexical form.

Structural transfer in Apertium is done by applying a set of rules in a left-to-right, longest-match fashion to prevent the translation from being performed word for word in those cases in which this would result in an incorrect translation. Structural transfer rules process sequences of lexical forms by performing operations such as reorderings or gender and number agreements. For the translation between non-related language pairs, such as Spanish–English, the structural transfer may be split into three levels in order to facilitate the writing of rules by linguists. The first level performs short-distance operations, such as gender and number agreement between nouns and adjectives, and groups sequences of lexical forms into *chunks*; second-level rules perform inter *chunk* operations, such as agreements between more distant constituents (i.e. subject and main verb); and third-level ones de-encapsulate the chunks and generate a sequence of TL lexical forms from each *chunk*. Note that, although the multi-level shallow transfer allows performing operations between words which are distant in the source sentence, shallow-transfer RBMT systems are less powerful that the ones which perform full parsing. In addition, each lexical form is processed at most by one rule in the same level.

The following example illustrates how lexical and structural transfer are performed in Apertium. Suppose that the Spanish sentence *Por otra parte mis amigos americanos han decidido venir* is to be translated into English. First, it is analysed as:

```
por otra parte<adv>
mío<det><pos><mf><pl>
amigo<n><m><pl>
americano<adj><m><pl>
haber<vbhaver><pri><p3><pl>
decidir<vblex><pp><m><sg>
venir<vblex><inf>
```

which splits the sentence in seven lexical forms: a multi-word adverb (*por otra parte*), a plural possessive determiner (*mío*), a noun and an adjective in masculine plural (*amigo* and *americano*, respectively), the third-person plural form of the present tense of the verb *to be* (*haber*), the masculine singular past participle of the verb *decidir* and the verb *venir* in infinitive mood. Then, the transfer step is executed. It starts by performing the lexical transfer and applying the first-level rules of the structural transfer in parallel. The lexical transfer of each SL lexical form gives as a result:

```
on the other hand<adv>
my<det><pos><pl>
friend<n><pl>
american<adj>
```

```
have<vbhaver><pres>
decide<vblex><pp>
come<vblex><inf>
```

Four first-level structural transfer rules are triggered: the first one matches a single adverb (the first lexical form in the example); the second one matches a determiner followed by an adjective and a noun (the next three lexical forms); the third one matches a form of the verb *haber* plus the past participle form of another verb (the next two lexical forms); and the last one matches a verb in infinitive mood (last lexical form). Each of these first-level rules group the matched lexical forms in the same *chunk* and perform local operations within the chunk; for instance, the second rule reorders the adjective and the noun:

```
ADV{ on the other hand<adv> }
NOUN_PHRASE{ my<det><pos><pl>
american<adj> friend<n><pl> }
HABER_PP{ have<vbhaver><pres>
decide<vblex><pp> }
INF{ come<vblex><inf> }
```

After that, inter *chunk* operations are performed. The *chunk* sequence *HABER_PP* (verb in present perfect tense) *INF* (verb in infinitive mood) matches a second-level rule which adds the preposition *to* between them:

```
ADV{ on the other hand<adv> }
NOUN_PHRASE{ my<det><pos><pl>
friend<n><pl> american<adj> }
HABER_PP{ have<vbhaver><pres>
decide<vblex><pp> }
TO{ to<pr> }
INF{ come<vblex><inf> }
```

Third-level structural transfer removes *chunk* encapsulations so that a plain sequence of lexical forms is generated:

```
on the other hand<adv>
my<det><pos><pl>
american<adj>
friend<n><pl>
have<vbhaver><pres>
decide<vblex><pp>
to<pr> come<vblex><inf>
```

Finally, the translation into TL is generated from the TL lexical forms: *On the other hand my American friends have decided to come*.

3 Related work

Linguistic data from RBMT have already been used to enrich SMT systems in different ways. Bilingual dictionaries have been added to SMT systems since its early days (Brown et al., 1993); one of the simplest strategies involves adding the dictionary entries directly to the training parallel corpus (Tyers, 2009; Schwenk et al., 2009). Other approaches go beyond that. Eisele et al. (2008) first translate the sentences in the test set with an RBMT system, then apply the usual phrase-extraction algorithm over the resulting small parallel corpus, and finally add the obtained phrase pairs to the original phrase table. It is worth noting that neither of these two strategies guarantee that the multi-word expressions in the RBMT bilingual dictionary appearing in the sentences to translate will be translated as such because they may be split into smaller units by the phrase-extraction algorithm. Our approach overcomes this issue by adding the data obtained from the RBMT system directly to the phrase table. Preliminary experiments with Apertium data shows that our hybrid approach outperforms the one by Eisele et al. (2008) when translating Spanish texts into English.

4 Enhancing phrase-based SMT with shallow-transfer linguistic resources

As already mentioned, the Apertium structural transfer detects sequences of lexical forms which need to be translated together to prevent them from being translated word for word, which would result in an incorrect translation. Therefore, adding to the phrase table of a PBSMT system all the bilingual phrase pairs which either match one of these sequences of lexical forms in the structural transfer or an entry in the bilingual dictionary suffices to encode all the linguistic information provided by Apertium. We add these bilingual phrase pairs directly to the phrase table, instead of adding them to the training corpus and rely on the phrase extraction algorithm (Koehn, 2010, sec. 5.2.3), to avoid splitting the multi-word expressions provided by Apertium into smaller phrases (Schwenk et al., 2009, sec. 2).

4.1 Phrase pair generation

Generating the set of bilingual phrase pairs which match bilingual dictionary entries is straightforward. First, all the SL surface forms that are recognised by Apertium and their corresponding lexical forms are generated. Then, these SL lexical forms are trans-

lated using the bilingual dictionary, and finally their TL surface forms are generated.

Bilingual phrase pairs which match structural transfer rules are generated in a similar way. First, the SL sentences to be translated are analysed to get their SL lexical forms, and then the sequences of lexical forms that either match a first-level or a second-level structural transfer rule are passed through the Apertium pipeline to get their translations. If a sequence of SL lexical forms is matched by more than one structural transfer rule in the same level, it will be used to generate as many bilingual phrase pairs as different rules it matches. This differs from the way in which Apertium translates, since in those case only the longest rule would be applied.

The following example illustrates this procedure. Let the Spanish sentence *Por otra parte mis amigos americanos han decidido venir*, from the example in the previous section, be one of the sentences to be translated. The SL sequences *por otra parte*, *mis amigos americanos*, *amigos americanos*, *han decidido*, *venir* and *han decidido venir* would be used to generate bilingual phrase pairs because they match a first-level rule, a second-level rule, or both. The SL words *amigos americanos* are used twice because they are covered by two first-level rules: one that matches a determiner followed by a noun and an adjective, and another that matches a noun followed by an adjective. Note that when using Apertium in the regular way, outside this hybrid approach, only the first rule is applied as a consequence of the left-to-right, longest match policy. The SL words *han decidido* and *venir* are used because they match first-level rules, whereas *han decidido venir* matches a second-level rule.

It is worth noting that the generation of bilingual phrase pairs from the shallow-transfer rules is guided by the test corpus. We decided to do it in this way in order to avoid meaningless phrases and also to make our approach computationally feasible. Consider, for instance, the rule which is triggered every time a determiner followed by a noun and an adjective is detected. Generating all the possible phrase pairs matching this rule would involve combining all the determiners in the dictionary with all the nouns and all the adjectives, causing the generation of many meaningless phrases, such as *el niño inalámbrico – the wireless boy*. In addition, the number of combinations to deal with becomes unmanageable as the length of the rule grows.

4.2 Scoring the new phrase pairs

State-of-the-art PBSMT systems usually attach 5 scores to every phrase pair in the translation table: source-to-target and target-to-source phrase translation probabilities, source-to-target and target-to-source lexical weightings, and phrase penalty.

To calculate the phrase translation probabilities of the phrase pairs obtained from the shallow-transfer RBMT resources we simply add them once to the list of corpus-extracted phrase pairs, and then compute the probabilities by relative frequency as it is usually done (Koehn, 2010, sec. 5.2.5). In this regard, it is worth noting that, as RBMT-generated phrase pairs are added only once, if one of them happens to share its source side with many other corpus-extracted phrase pairs, or even with a single, very frequent one, the RBMT-generated phrase pair will receive lower scores, which penalises its use. To alleviate this without adding the same phrase pair an arbitrary amount of times, we introduce an additional boolean score to flag phrase pairs obtained from the RBMT resources.

The fact that the generation of bilingual phrase pairs from shallow transfer rules is guided by the test corpus may cause the translation of a sentence to be influenced by other sentences in the test set. This happens when the translation provided by Apertium for a subsegment of a test sentence matching an Apertium structural transfer rule is shared with one or more subsegments in the test corpus. In that case, the phrase translation probability $p(\mathrm{source}|\mathrm{target})$ of the resulting bilingual phrase pair is lower than if no subsegments with the same translation were found.

To calculate the lexical weightings (Koehn, 2010, sec. 5.3.3) of the RBMT-generated phrase pairs, the alignments between the words in the source side and those in the target side are needed. These word alignments are obtained by tracing back the operations carried out in the different steps of the shallow-transfer RBMT system. Only those words which are neither split nor joint with other words by the RBMT engine are included in the alignments; thus, multi-word expressions are left unaligned. This is done for convenience, since in this way multi-word

Figure 1: Example of word alignment obtained by tracing back the operations done by Apertium when translating from Spanish to English the sentence *Por otra parte mis amigos americanos han decidido venir*. Note that *por otra parte* is analysed by Apertium as a multi-word expression whose words are left unaligned for convenience (see section 4.2).

expressions are assigned a lexical weighting of 1.0. Figure 1 shows the alignment between the words in the running example.

5 System training

We submitted a hybrid system for the Spanish–English language pair built by following the strategy described above. The initial phrase table was built from all the parallel corpora distributed as part of the WMT 2011 shared translation task, namely Europarl (Koehn, 2005), News Commentary and United Nations. In a similar way, the language model was built from the the Europarl (Koehn, 2005) and the News Crawl monolingual English corpora. The weights of the different feature functions were optimised by means of minimum error rate training (Och, 2003) on the 2008 test set.[1] Table 1 summarises the data about the corpora used to build our submission. We also built a baseline PBSMT system trained on the same corpora and a reduced version of our system whose phrase table was enriched only with dictionary entries.

The Apertium (Forcada et al., 2011) engine and the linguistic resources for Spanish–English were downloaded from the Apertium Subversion repository.The linguistic data contains 326 228 entries in the bilingual dictionary, 106 first-level structural transfer rules, and 31 second-level rules. As entries in the bilingual dictionary contain mappings between SL and TL lemmas, when phrase pairs matching the bilingual dictionary are generated all the possible inflections of these lemmas are produced.

We used the free/open-source PBSMT system Moses (Koehn et al., 2007), together with the IRSTLM language modelling toolkit (Federico et al., 2008), which was used to train a 5-gram language model using interpolated Kneser-Ney discounting (Goodman and Chen, 1998). Word alignments from the training parallel corpus were computed by means of GIZA++ (Och and Ney, 2003). The cube pruning (Huang and Chiang, 2007) decoding algorithm was chosen in order to speed-up the tuning step and the translation of the test set.

Task	Corpus	Sentences
Language model	Europarl	2 015 440
	News Crawl	112 905 708
	Total	114 921 148
Training	Europarl	1 786 594
	News Commentary	132 571
	United Nations	10 662 993
	Total	12 582 158
	Total clean	8 992 751
Tuning	newstest2008	2 051
Test	newstest2011	3 003

Table 1: Size of the corpora used in the experiments. The bilingual training corpora has been cleaned to remove empty parallel sentences and those which contain more than 40 tokens.

6 Results and discussion

Table 2 reports the translation performance as measured by BLEU (Papineni et al., 2002), GTM (Melamed et al., 2003) and METEOR[2] (Banerjee and Lavie, 2005) for Apertium and the three systems presented in the previous section, as well as the size of the phrase table and the amount of unknown words in the test set. The hybrid approach outperforms the baseline PBSMT system in terms of the three evaluation metrics. The confidence interval of the difference between them, computed by doing 1 000 iterations of paired

[1]The corpora can be downloaded from `http://www.statmt.org/wmt11/translation-task.html`.

[2]Modules *exact*, *stem*, *synonym* and *paraphrase* (Denkowski and Lavie, 2010) were used.

system	BLEU	GTM	METEOR	# of unknown words	phrase table size
baseline	28.06	52.40	47.27	1 447	254 693 494
UA-dict	28.58	52.55	47.41	1 274	255 860 346
UA	**28.73**	**52.66**	**47.51**	1 274	255 872 094
Apertium	23.89	50.71	45.65	4 064	-

Table 2: Case-insensitive BLEU, GTM, and METEOR scores obtained by the hybrid approach submitted to the WMT 2011 shared translation task (*UA*), a reduced version of it whose phrase table is enriched using only bilingual dictionary entries (*UA-dict*), a baseline PBSMT system trained with the same corpus (*baseline*), and Apertium on the *newstest2011* test set. The number of unknown words and the phrase table size are also reported when applicable.

bootstrap resampling (Zhang et al., 2004) with a p-level of 0.05, does not overlap with zero for any evaluation metric,[3] which confirms that it is statistically significant. Our hybrid approach also outperforms Apertium in terms of the three evaluation metrics.[4] However, the difference between our complete hybrid system and the version which only takes advantage of bilingual dictionary is not statistically significant for any metric.[5]

The results show how the addition of RBMT-generated data leads to an improvement over the baseline PBMST system, even though it was trained with a very large parallel corpus and the proportion of entries from the Apertium data in the phrase table is very small (0.46%). 5.94% of the phrase pairs chosen by the decoder were generated from the Apertium data. The improvement may be explained by the fact that the sentences in the test set belong to the news domain and Apertium data has been developed bearing in mind the translation of general texts (mainly news), whereas most of the bilingual training corpus comes from specialised domains. In addition, the morphology of Spanish is quite rich, which makes it very difficult to find all possible inflections of the same lemma in a parallel corpus. Therefore, Apertium-generated phrases, which contain handcrafted knowledge from a general domain, cover some sequences of words in the input text which are not covered, or are sparsely found, in the original training corpora, as shown by the reduction in the amount of unknown words (1 447 unknown words versus 1 274). In other words, Apertium linguistic information does not completely overlap with the data learned from the parallel corpus. Regarding the small difference between the hybrid system enriched with all the Apertium resources and the one that only includes the bilingual dictionary, preliminary experiments shows that the impact of the shallow-transfer rules is higher when the TL is highly inflected and the SL is not, which is exactly the scenario opposite to the one described in this paper.

[3]The confidence interval of the difference between our system and the baseline PBSMT system for BLEU, GTM and METEOR is $[0.38, 0.93]$, $[0.06, 0.45]$, and $[0.06, 0.42]$, respectively.

[4]The confidence interval of the difference between our approach and Apertium for BLEU, GTM and METEOR is $[4.35, 5.35]$, $[1.55, 2.32]$, and $[1.50, 2.21]$, respectively.

[5]The confidence interval of the difference between our approach and the reduced version which does not use structural transfer rules for BLEU, GTM and METEOR is $[-0.07, 0.37]$, $[-0.06, 0.27]$, and $[-0.06, 0.26]$, respectively.

7 Concluding remarks

We have presented the MT system submitted by the Transducens Research Group from Universitat d'Alacant to the WMT2011 shared translation task. This is the first submission of our team to this shared task. We developed a hybrid system for the Spanish–English language pair which enriches the phrase table of a standard PBSMT system with phrase pairs generated from the RBMT linguistic resources provided by Apertium. Our system outperforms a baseline PBSMT in terms of BLEU, GTM and METEOR scores by a statistically significant margin.

Acknowledgements

Work funded by the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01 and by Generalitat Valenciana through grant ACIF/2010/174 (VALi+d programme).

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, M. J. Goldsmith, J. Hajic, R. L. Mercer, and S. Mohanty. 1993. But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, pages 202–205, Princeton, New Jersey.

M. Denkowski and A. Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*, pages 339–342, Uppsala, Sweden.

A. Eisele, C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio.

M. Federico, N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH-2008*, pages 1618–1621, Brisbane, Australia.

M.L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J.A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F.M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation. Special Issue on Free/Open-Source Machine Translation*, In press.

J. Goodman and S. F. Chen. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, August.

L. Huang and D. Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic.

W. J. Hutchins and H. L. Somers. 1992. *An introduction to machine translation*, volume 362. Academic Press New York.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 48–54, Edmonton, Canada.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, C. Shen, W.and 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 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5:12–16.

P. Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

I. D. Melamed, R. Green, and J. P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 61–63, Edmonton, Canada.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March.

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

K. Papineni, S. Roukos, T. Ward, and W. 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, Philadelphia, Pennsylvania.

H. Schwenk, S. Abdul-Rauf, L. Barrault, and J. Senellart. 2009. SMT and SPE machine translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134, Athens, Greece.

G. Thurmair. 2009. Comparing different architectures of hybrid Machine Translation systems. In *Proceedings MT Summit XII*, Ottawa, Ontario, Canada.

F. M. Tyers. 2009. Rule-based augmentation of training data in Breton-French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 213–217, Barcelona, Spain.

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2051–2054, Lisbon, Portugal.