Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics

Chi-kiu LO and **Dekai WU**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo,dekai}@cs.ust.hk

Abstract

We present an unsupervised approach to estimate the appropriate degree of contribution of each semantic role type for semantic translation evaluation, yielding a semantic MT evaluation metric whose correlation with human adequacy judgments is comparable to that of recent supervised approaches but without the high cost of a human-ranked training corpus. Our new unsupervised estimation approach is motivated by an analysis showing that the weights learned from supervised training are distributed in a similar fashion to the relative frequencies of the semantic roles. Empirical results show that even without a training corpus of human adequacy rankings against which to optimize correlation, using instead our relative frequency weighting scheme to approximate the importance of each semantic role type leads to a semantic MT evaluation metric that correlates comparable with human adequacy judgments to previous metrics that require far more expensive human rankings of adequacy over a training corpus. As a result, the cost of semantic MT evaluation is greatly reduced.

1 Introduction

In this paper we investigate an unsupervised approach to estimate the degree of contribution of each semantic role type in semantic translation evaluation in low cost without using a human-ranked training corpus but still yields a evaluation metric that correlates comparably with human adequacy judgments to that of recent supervised approaches as in Lo and Wu (2011a, b, c). The new approach is motivated by an analysis showing that the distribution of the weights learned from the supervised training is similar to the relative frequencies of the occurrences of each semantic role in the reference translation. We then introduce a relative frequency weighting scheme to approximate the importance of each semantic role type. With such simple weighting scheme, the cost of evaluating translation of languages with fewer resources available is greatly reduced.

For the past decade, the task of measuring the performance of MT systems has relied heavily on lexical n-gram based MT evaluation metrics, such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), PER (Tillmann *et al.*, 1997), CDER (Leusch *et al.*, 2006) and WER (Nießen *et al.*, 2000) because of their support on fast and inexpensive evaluation. These metrics are good at ranking overall systems by averaging their scores over the entire document. As MT systems improve, the focus of MT evaluation changes from generally reflecting the quality of each system to assisting error analysis on each MT output in detail. The failure of such metrics in evaluating translation quality on sentence level are becoming more apparent. Though containing roughly the correct words, the MT output as a whole sentence is still quite incomprehensible and fails to express meaning that is close to the input. Lexical n-gram based evaluation metrics are surface-oriented and do not do so well at ranking translations according to adequacy and are particularly poor at reflecting significant translation quality improvements on more meaningful word sense or semantic frame choices which human judges can indicate clearly. Callison-Burch *et al.* (2006) and Koehn and Monz (2006) even reported cases where BLEU strongly disagrees with human judgment on translation quality.

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 49–56,
Jeju, Republic of Korea, 12 July 2012.

Liu and Gildea (2005) proposed STM, a structural approach based on syntax to addresses the failure of lexical similarity based metrics in evaluating translation grammaticality. However, a grammatical translation can achieve a high syntax-based score but still contains meaning errors arising from confusion of semantic roles. On the other hand, despite the fact that non-automatic, manually evaluations, such as HTER (Snover *et al.*, 2006), are more adequacy oriented and show a high correlation with human adequacy judgment, the high labor cost prohibits their widespread use. There was also work on explicitly evaluating MT adequacy with aggregated linguistic features (Giménez and Màrquez, 2007, 2008) and textual entailment (Pado *et al.*, 2009).

In the work of Lo and Wu (2011a), MEANT and its human variants HMEANT were introduced and empirical experimental results showed that HMEANT, which can be driven by low-cost monolingual semantic roles annotators with high inter-annotator agreement, correlates as well as HTER and far superior than BLEU and other surfaced oriented evaluation metrics. Along with additional improvements to the MEANT family of metrics, Lo and Wu (2011b) detailed the studies of the impact of each individual semantic role to the metric's correlation with human adequacy judgments. Lo and Wu (2011c) further discussed that with a proper weighting scheme of semantic frame in a sentence, structured semantic role representation is more accurate and intuitive than flattened role representation for semantic MT evaluation metrics.

The recent trend of incorporating more linguistic features into MT evaluation metrics raise the discussion on the appropriate approach in weighting and combining them. ULC (Giménez and Màrquez, 2007, 2008) uses uniform weights to aggregate linguistic features. This approach does not capture the importance of each feature to the overall translation quality to the MT output. One obvious example of different semantic roles contribute differently to the overall meaning is that readers usually accept translations with errors in adjunct arguments as a valid translation but not those with errors in core arguments. Unlike ULC, Liu and Gildea (2007); Lo and Wu (2011a) approach the weight estimation problem by maximum correlation training which directly optimize the correlation with human adequacy judgments. However, the shortcomings of this approach is that it requires a human-ranked training corpus which is expensive, especially for languages with limited resource.

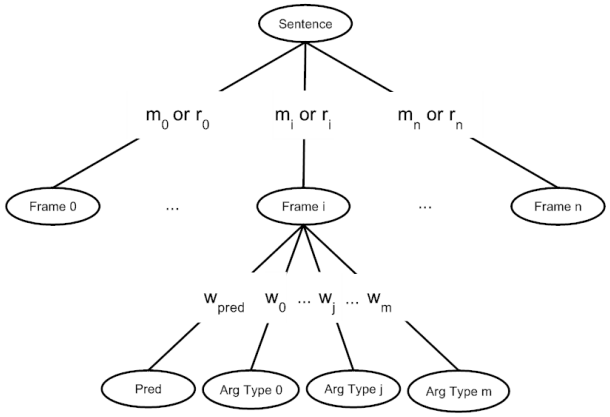

Figure 1: HMEANT structured role representation with a weighting scheme reflecting the degree of contribution of each semantic role type to the semantic frame. (Lo and Wu, 2011a,b,c).

We argue in this paper that for semantic MT evaluation, the importance of each semantic role type can easily be estimated using a simple unsupervised approach which leverage the relative frequencies of the semantic roles appeared in the reference translation. Our proposed weighting scheme is motivated by an analysis showing that the weights learned from supervised training are distributed in a similar fashion to the relative frequencies of the semantic roles. Our results show that the semantic MT evaluation metric using the relative frequency weighting scheme to approximate the importance of each semantic role type correlates comparably with human adequacy judgments to previous metrics that use maximum correlation training, which requires expensive human rankings of adequacy over a training corpus. Therefore, the cost of semantic MT evaluation is greatly reduced.

2 Semantic MT evaluation metrics

Adopting the principle that a good translation is one from which human readers may successfully understand at least the basic event structure-"who did what to whom, when, where and why" (Pradhan *et al.*, 2004)-which represents the most essential meaning of the source utterances, Lo and Wu (2011a,b,c)

proposed HMEANT to evaluate translation utility based on semantic frames reconstructed by human reader of machine translation output. Monolingual (or bilingual) annotators must label the semantic roles in both the reference and machine translations, and then to align the semantic predicates and role fillers in the MT output to the reference translations. These annotations allow HMEANT to then look at the aligned role fillers, and aggregate the translation accuracy for each role. In the spirit of Occam's razor and representational transparency, the HMEANT score is defined simply in terms of a weighted f-score over these aligned predicates and role fillers. More precisely, HMEANT is defined as follows:

1. Human annotators annotate the shallow semantic structures of both the references and MT output.

2. Human judges align the semantic frames between the references and MT output by judging the correctness of the predicates.

3. For each pair of aligned semantic frames,

   (a) Human judges determine the translation correctness of the semantic role fillers.

   (b) Human judges align the semantic role fillers between the reference and MT output according to the correctness of the semantic role fillers.

4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers.

$$m_i \equiv \frac{\text{\#tokens filled in frame i of MT}}{\text{total \#tokens in MT}}$$
$$r_i \equiv \frac{\text{\#tokens filled in frame i of REF}}{\text{total \#tokens in REF}}$$

$$M_{i,j} \equiv \text{total \# ARG j of PRED i in MT}$$
$$R_{i,j} \equiv \text{total \# ARG j of PRED i in REF}$$
$$C_{i,j} \equiv \text{\# correct ARG j of PRED i in MT}$$
$$P_{i,j} \equiv \text{\# partially correct ARG j of PRED i in MT}$$

$$\text{precision} = \frac{\sum_i m_i \frac{w_{\text{pred}} + \sum_j w_j (C_{i,j} + w_{\text{partial}} P_{i,j})}{w_{\text{pred}} + \sum_j w_j M_{i,j}}}{\sum_i m_i}$$

$$\text{recall} = \frac{\sum_i r_i \frac{w_{\text{pred}} + \sum_j w_j (C_{i,j} + w_{\text{partial}} P_{i,j})}{w_{\text{pred}} + \sum_j w_j R_{i,j}}}{\sum_i r_i}$$

$$\text{HMEANT} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

where m_{i} and r_{i} are the weights for frame,i, in the MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. $M_{i,j}$ and $R_{i,j}$ are the total counts of argument of type j in frame i in the MT/REF respectively. $C_{i,j}$ and $P_{i,j}$ are the count of the correctly and partial correctly translated argument of type j in frame i in the MT. w_{pred} is the weight for the predicate and w_{j} is the weights for the arguments of type j. These weights estimate the degree of contribution of different types of semantic roles to the overall meaning of the semantic frame they attached to. The frame precision/recall is the weighted sum of the number of correctly translated roles in a frame normalized by the weighted sum of the total number of all roles in that frame in the MT/REF respectively. The sentence precision/recall is the weighted sum of the frame precision/recall for all frames normalized by the weighted sum of the total number of frames in MT/REF respectively. Figure 1 shows the internal structure of HMEANT.

In the work of Lo and Wu (2011b), the correlation of all individual roles with the human adequacy judgments were found to be non-negative. Therefore, grid search was used to estimate the weights of each roles by optimizing the correlation with human adequacy judgments. This approach requires an expensive human-ranked training corpus which may not be available for languages with sparse resources.Unlike the supervised training approach, our proposed relative frequency weighting scheme does not require additional resource other than the SRL annotated reference translation.

3 Which roles contribute more in the semantic MT evaluation metric?

We begin with an investigation that suggests that the relative frequency of each semantic role (which can be estimated in unsupervised fashion without human rankings) approximates fairly closely its importance as determined by previous supervised optimization approaches. Since there is no ground truth on which

Role	Deviation (GALE-A)	Deviation (GALE-B)	Deviation (WMT12)
Agent	-0.09	-0.05	0.03
Experiencer	0.23	0.05	0.02
Benefactive	0.02	0.04	-0.01
Temporal	0.11	0.08	0.03
Locative	-0.05	-0.05	-0.07
Purpose	-0.01	0.03	-0.01
Manner	-0.01	0.00	-0.01
Extent	-0.02	0.00	-0.01
Modal	—	0.04	0.01
Negation	—	0.01	-0.01
Other	-0.12	0.05	-0.01

Table 1: Deviation of relative frequency from optimized weight of each semantic role in GALE-A, GALE-B and WMT12

semantic role contribute more to the overall meaning in a sentence for semantic MT evaluation, we first show that the unsupervised estimation are close to the weights obtained from the supervised maximum correlation training on a human-ranked MT evaluation corpus. More precisely, the weight estimation function is defined as follows:

$$c_j \equiv \text{\# count of ARG j in REF of the test set}$$

$$w_j = \frac{c_j}{\sum_j c_j}$$

3.1 Experimental setup

For our benchmark comparison, the evaluation data for our experiment is the same two sets of sentences, GALE-A and GALE-B that were used in Lo and Wu (2011b). The translation in GALE-A is SRL annotated with 9 semantic role types, while those in GALE-B are SRL annotated with 11 semantic role types (segregating the *modal* and the *negation* roles from the *other* role).

To validate whether or not our hypothesis is language independent, we also construct an evaluation data set by randomly selecting 50 sentences from WMT12 English to Czech (WMT12) translation task test corpus, in which 5 systems (out of 13 participating systems) were randomly picked for translation adequacy ranking by human readers. In total, 85 sets of translations (with translations from some source sentences appear more than once in different sets) were ranked. The translation in WMT12 are also SRL annotated with the tag set as GALE-B, i.e., 11 semantic role types.

The weights w_{pred}, w_j and w_{partial} were estimated using grid search to optimize the correlation against human adequacy judgments.

3.2 Results

Inspecting the distribution of the trained weights and the relative frequencies from all three data sets, as shown in table 1, we see that the overall pattern of weights from unsupervised estimation has a fairly small deviation from the those learned via supervised optimization. To visualize more clearly the overall pattern of the weights from the two estimation methods, we show the deviation of the unsupervised estimation from the supervised estimation. A deviation of 0 for all roles would mean that unsupervised and supervised estimation produce exactly identical weights. If the unsupervised estimation is higher than the supervised estimation, the deviation will be positive and vice versa.

What we see is that in almost all cases, the deviation between the trained weight and the relative frequency of each role is always within the range [-0.1, 0.1].

Closer inspection also reveals the following more detailed patterns:

- The weight of the less frequent adjunct arguments (e.g. purpose, manner, extent, modal and negation) from the unsupervised estimation is highly similar to that learned from the super-

PRED estimation	Deviation (GALE-A)	Deviation (GALE-B)	Deviation (WMT12)
Method (i)	0.16	0.16	0.31
Method (ii)	0.02	0.01	0.01

Table 2: Deviation from optimized weight in GALE-A, GALE-B and WMT12 of the predicate's weight as estimated by (i) frequency of predicates in frames, relative to predicates and arguments; and (ii) one-fourth of agent's weight.

vised maximum correlation training.

- The unsupervised estimation usually gives a higher weight to the temporal role than the supervised training would.
- The unsupervised estimation usually gives a lower weight to the locative role than the supervised training would but the two weights from the two approach are still high similar to each other, yielding a deviation within the range of [-0.07, 0.07].
- There is an obvious outlier found in GALE-A where the deviation of the relative frequency from the optimized weight is unusually high. This suggests that the optimized weights in GALE-A may be at the risk of over-fitting the training data.

4 Estimating the weight for the predicate

The remaining question left to be investigated is how we are to estimate the importance of the predicate in an unsupervised approach. One obvious approach is to treat the predicate the same way as the arguments. That is, just like with arguments, we could weight predicates by the relative frequency of how often predicates occur in semantic frames. However, this does not seem well motivated since predicates are fundamentally different from arguments: by definition, every semantic frame is defined by one predicate, and arguments are defined relative to the predicate.

On the other hand, inspecting the weights on the predicate obtained from the supervised maximum correlation training, we find that the weight of the predicate is usually around one-fourth of the weight of the agent role. More precisely, the two weight estimation functions are defined as follows:

$$c_{\text{pred}} \equiv \text{\# count of PRED in REF of the test set}$$

$$\begin{aligned} \text{Method (i)} &= \frac{c_{\text{pred}}}{c_{\text{pred}} + \sum_j c_j} \\ \text{Method (ii)} &= 0.25 \cdot w_{\text{agent}} \end{aligned}$$

We now show that the supervised estimation of the predicate's weight is closely approximated by unsupervised estimation.

4.1 Experimental setup

The experimental setup is the same as that used in section 3.

4.2 Results

The results in table 2 show that the trained weight of the predicate and its unsupervised estimation of one-fourth of the agent role's weight are highly similar to each other. In all three data sets, the deviation between the trained weight and the heuristic of one-fourth of the agent's weight is always within the range [0.1, 0.2].

On the other hand, treating the predicate the same as arguments by estimating the unsupervised weight using relative frequency largely over-estimates and has a large deviation from the weight learned from supervised estimation.

5 Semantic MT evaluation using unsupervised weight estimates

Having seen that the weights of the predicate and semantic roles estimated by the unsupervised approach fairly closely approximate those learned from the supervised approach, we now show that the unsupervised approach leads to a semantic MT evaluation metric that correlates comparably with human adequacy judgments to one that is trained on a far more expensive human-ranked training corpus.

5.1 Experimental setup

Following the benchmark assessment in NIST MetricsMaTr 2010 (Callison-Burch *et al.*, 2010), we assess the performance of the semantic MT evaluation

Metrics	GALE-A	GALE-B	WMT12
HMEANT (supervised)	0.49	0.27	0.29
HMEANT (unsupervised)	0.42	0.23	0.20
NIST	0.29	0.09	0.12
METEOR	0.20	0.21	0.22
TER	0.20	0.10	0.12
PER	0.20	0.07	0.02
BLEU	0.20	0.12	0.01
CDER	0.12	0.10	0.14
WER	0.10	0.11	0.17

Table 3: Average sentence-level correlation with human adequacy judgments of HMEANT using supervised and unsupervised weight scheme on GALE-A, GALE-B and WMT12, (with baseline comparison of commonly used automatic MT evaluation metric.

metric at the sentence level using Kendall's rank correlation coefficient which evaluate the correlation of the proposed metric with human judgments on translation adequacy ranking. A higher the value for indicates a higher similarity to the ranking by the evaluation metric to the human judgment. The range of possible values of correlation coefficient is [-1,1], where 1 means the systems are ranked in the same order as the human judgment and -1 means the systems are ranked in the reverse order as the human judgment. For GALE-A and GALE-B, the human judgment on adequacy was obtained by showing all three MT outputs together with the Chinese source input to a human reader. The human reader was instructed to order the sentences from the three MT systems according to the accuracy of meaning in the translations. For WMT12, the human adequacy judgments are provided by the organizers.

The rest of the experimental setup is the same as that used in section 3.

5.2 Results

Table 3 shows that HMEANT with the proposed unsupervised semantic role weighting scheme correlate comparably with human adequacy judgments to that optimized with a more expensive human-ranked training corpus, and, outperforms all other commonly used automatic metrics (except for METEOR in Czech). The results from GALE-A, GALE-B and WMT12 are consistent. These encouraging results show that semantic MT evaluation metric could be widely applicable to languages other than English.

6 Conclusion

We presented a simple, easy to implement yet well-motivated weighting scheme for HMEANT to estimate the importance of each semantic role in evaluating the translation adequacy. Unlike the previous metrics, the proposed metric does not require an expensive human-ranked training corpus and still outperforms all other commonly used automatic MT evaluation metrics. Interestingly, the distribution of the optimal weights obtained by maximum correlation training, is similar to the relative frequency of occurrence of each semantic role type in the reference translation. HMEANT with the new weighting scheme showed consistent results across different language pairs and across different corpora in the same language pair. With the proposed weighting scheme, the semantic MT evaluation metric is ready to be used off-the-shelf without depending on a human-ranked training corpus. We believe that our current work reduces the barrier for semantic MT evaluation for resource scarce languages sufficiently so that semantic MT evaluation can be applied to most other languages.

Acknowledgments

We would like to thank Ondřej Bojar and all the annotators from the Charles University in Prague for participating in the experiments. This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the Eu-

ropean Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF621008, GRF612806, DAG03/04.EG09, RGC6256/00E, and RGC6083/99E. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the RGC, EU, or DARPA.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 65–72, 2005.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in Machine Translation Research. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 249–256, 2006.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Pryzbocki, and Omar Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, 15-16 July 2010.

G. Doddington. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT-02)*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Jesús Giménez and Lluís Màrquez. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 195–198, Columbus, OH, June 2008. Association for Computational Linguistics.

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

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 2006.

Ding Liu and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 25, 2005.

Ding Liu and Daniel Gildea. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of the 2007 Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-07)*, 2007.

Chi-kiu Lo and Dekai Wu. MEANT: An Inexpensive, High-Accuracy, Semi-Automatic Metric for Evaluating Translation Utility based on Semantic Roles. In *Proceedings of the Joint conference of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT-11)*, 2011.

Chi-kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

Chi-kiu Lo and Dekai Wu. Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation. In *Proceedings of the 5th Workshop on Syntax and Structure in Statistical Translation (SSST-5)*, 2011.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, 2000.

Sebastian Pado, Michel Galley, Dan Jurafsky, and Chris Manning. Robust Machine Translation Evaluation with Entailment Features. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP-09)*, 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, 2002.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of*

the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04), 2004.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–231, 2006.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. Accelerated DP Based Search For Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97)*, 1997.