# A semantically confidence-weighted ITG induction algorithm

Meriem Beloucif and Dekai Wu
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{mbeloucif|dekai}@cs.ust.hk

## Abstract

We propose a new algorithm to induce inversion transduction grammars, in which a crosslingual semantic frame based objective function is injected as confidence weighting in the early stages of statistical machine translation training. Unlike recent work on improving translation adequacy that uses a monolingual semantic frame based objective function to drive the tuning of loglinear mixture weights in the late stages of statistical machine translation training, our bilingual approach incorporates the semantic objective during the actual learning of the translation model's structure. Our approach assigns higher confidence to training examples in which the semantic frames in the input language more closely match the semantic frames of the output language, as predicted automatically by XMEANT, the crosslingual semantic frame based machine translation evaluation metric. We chose to apply this approach to induce inversion transduction grammars (ITGs), since ITG alignments prune a large majority of the space of possible alignments, while at the same time empirically fully covering all the crosslingual semantic frame alternations of the type we are using for confidence weighting. Results show that boosting semantically compatible training examples in ITG induction improves the translation performance compared to either traditional GIZA++ alignment or conventional ITG alignment based approaches for phrase based statistical machine translation.



In Proceedings of 3rd International Workshop on Semantic Machine Learning (SML 2016), 10th July 2016, New York City, NY, USA.

## 1 Introduction

In this paper we introduce an approach that uses a semantic based objective function at a very early stage of training statistical machine translation (SMT) systems, more precisely, during the actual learning of the translation model's structure. Recent research has shown that including a semantic based objective function in the training pipeline, such as tuning against semantic based metrics like MEANT (Lo *et al.*, 2012), improves the translation adequacy (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b; Beloucif *et al.*, 2014). We show that integrating a semantic based objective function much earlier in the training produces a more semantically correct alignment. Our approach is also motivated by the fact that XMEANT (Lo *et al.*, 2014), a crosslingual semantic evaluation metric, has been shown to correlate better with human adequacy judgement than most commonly used evaluation metrics under some conditions. Our algorithm assigns a higher confidence to training examples in which XMEANT performs well, in other words, for bisentences where the semantic frames in the input language match more closely the semantic frames of the output language. We also show that this way of inducing ITGs does not only improve the translation quality, but it also produces better alignments in comparison to conventional ITG alignments and to the traditional GIZA++ (Och and Ney, 2000) alignments.

Applying this approach to induce inversion transduction grammars is also motivated by the fact that ITG alignments have previously been empirically shown to cover almost all crosslingual semantic frame alternations, even though they rule out the majority of incorrect alignments (Addanki *et al.*, 2012). We show that using a confidence-weighting algorithm for ITG induction not only helps further narrow down the ITGs constraints even more, but also avoids losing relevant portions of the search space, thus learning a more semantically driven word alignment. We deliberately train our approach using a relatively small data set to show that a semantic based learning can also help a lot with low resource languages in comparison to existing learning methods. Although Chinese is not a low resource language, we are deliberately simulating low resource conditions in our experiments by training on a relatively small parallel data set.

## 2 Related work

### 2.1 Crosslingual evaluation metric XMEANT

Our approach implements the principle that a good translation is one where a human can easily understand the general meaning of the output sentence as captured by the basic event structure: *who did what to whom, when, where and why* as defined by Pradhan *et al.* (2004). The MEANT family of metrics are semantic evaluation metrics that have been shown to correlate more closely with human adequacy judgement than the most commonly used surface based metrics (Lo and Wu, 2011, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b; Macháček and Bojar, 2013). MEANT compares the MT output sentence against the provided reference translations, and produces a score to measure the degree of similarity between their semantic frame structures. Our new approach is encouraged by the fact that many previous studies have empirically shown that integrating semantic role labeling into the training pipeline by tuning against MEANT improves the translation adequacy (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b; Beloucif *et al.*, 2014)

XMEANT (Lo *et al.*, 2014) is a crosslingual version of the semantic evaluation metric MEANT. It has been shown in some cases to correlate even better with human adequacy judgments than MEANT, and also better than most evaluation metrics like BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006).

Unlike MEANT which requires expensive man-made reference translations, XMEANT uses semantically parsed foreign language input to evaluate the MT translation output. MEANT measures lexical similarity using a monolingual context vector model, whereas XMEANT substitutes simple crosslingual lexical translation probabilities. Figure 1 describes the XMEANT algorithm. Each token of the role fillers in the input string is aligned to the token of the role fillers in the output string that has the maximum lexical translation probability. XMEANT crosslingual phrasal similarities are computed as follows (Lo *et al.*, 2014) :

**Algorithm** XMEANT

1. Apply *an input language* automatic shallow semantic parser *to the foreign input* and *an output language* automatic shallow semantic parser *to the MT output*.
2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between *the foreign input* and *the MT output* according to **the lexical translation probabilities** of the predicates.
3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between *the foreign input* and *the MT output* according to the **aggregated phrasal translation probabilities** of the role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the definitions similar to MEANT.

Figure 1: XMEANT algorithm

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i \\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i \\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i \\
p(e,f) &= \sqrt{t\left(e|f\right)t\left(f|e\right)} \\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max\limits_{f\in\mathbf{f}} p(e,f)}{|\mathbf{e}|} \\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max\limits_{e\in\mathbf{e}} p(e,f)}{|\mathbf{f}|} \\
s_{i,\text{pred}} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}\cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}+\text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}} \\
s_{i,j} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}\cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}+\text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

where the joint probability $p$ is the harmonic mean of the two directions of the translation table $t$ trained using IBM model 1 (Brown *et al.*, 1993). $\text{prec}_{\mathbf{e},\mathbf{f}}$ is the precision and $\text{rec}_{\mathbf{e},\mathbf{f}}$ is the recall of the phrasal similarities of the role fillers. $s_{i,\text{pred}}$ and $s_{i,j}$ are the f-scores of the phrasal similarities of the predicates and role fillers of the arguments of type $j$ between the input and the MT output.

Our approach takes advantage of the fact that XMEANT judges the extent to which the semantic frames of the input match those of the output, using these highly informative scores as confidence weights for training examples. We show that by using this convenient method to inject a crosslingual semantic reward/error signal into the ITG induction algorithm enables us not only to learn more semantically based correlations between the two languages, but also that this semantic bias helps under low resource conditions.

### 2.2 Alignment

Conventional alignment algorithms such as IBM models (Brown *et al.*, 1990) and HMM models (Vogel *et al.*, 1996) are flat and directed. They need two separate asymmetric alignments to form a single *bidirectional alignment*, then use heuristics to harmonize the two directed alignments, as implemented in GIZA++ (Och and Ney, 2000). This means that there is no model that considers the final bidirectional alignment where the translation system is trained on to be optimal. Transduction grammars (Wu, 1997), on the other hand, have proven that learning word alignments using a system that is compositionally structured can provide *optimal bidirectional alignments.* Although this structured optimality comes at a higher cost in terms of time complexity, it allows preexisting structured information to be incorporated into the model.

The generative capacity of ITGs puts in place efficient and universal hypothesis language translation constraints. The ITG hypothesis assumes that sentence translations between any two languages can be accomplished within the expressivenes of the ITG formalism which results in learning generalizations over bilingual relations without exploding the model complexity. Saers and Wu (2009) proposed a better method of producing word alignment by training inversion transduction grammars (Wu, 1997). One problem encountered with such model was the complexity of the biparsing algorithm which runs in $O(n^6)$ . A faster algorithm that runs in $O(n^3)$ (Saers *et al.*, 2010) was proposed later. Zhang and Gildea (2005) presented a version of ITG where rule probabilities are lexicalized throughout the synchronous parse tree for efficient training which helped align sentences up to 15 words.

Some of the previous work on word alignment used morphological and syntactic features (De Gispert *et al.*, 2006). Some log linear models have been proposed to incorporate those features (Dyer *et al.*, 2011). The problem with these approaches is that they require language specific knowledge and that they always work better on more morphologically rich languages. A few studies that approximately integrate semantic knowledge in computing word alignment are proposed by Ma *et al.* (2011) and Songyot and Chiang (2014). However, the former needs to have a prior word alignment learned on lexical words. The authors in the latter model proposed a semantic oriented word alignment. However, word similarities first need to be extracted from monolingual data, and are then used to produce alignments.

## 3 Confidence-weighted training algorithm

We implemented a token based bracketing inversion transduction grammars (BITG) as our ITG system. BITGs have been proven to produce a good result by only using one nonterminal category (Saers *et al.*, 2009). The algorithm we propose in this paper uses the crosslingual semantic evaluation metric XMEANT as a confidence weighting metric in the early stages of statistical machine translation training. We modify the BITG induction algorithm of Saers *et al.* (2009), weighting training examples using the confidence as judged by XMEANT, i.e., we weight training examples according to how closely the semantic frames in the input language match the semantic frames of the output language semantic frame. In this way we are biasing the bracketing inversion transduction grammar (BITG) towards preferring bilingual parses that better fit XMEANT's crosslingual semantic frames.

We contrast our new proposed model to the token based BITG system. We initialize both ITG based models with uniform structural probabilities, setting aside half of the probability mass for lexical rules. This probability mass is distributed among the lexical rules according to co-occurrence counts from the training data, assuming each sentence to contain one empty token to account for singletons. These initial probabilities are refined with 10 iterations of expectation maximization where the expectation step is calculated using beam pruned parsing (Saers *et al.*, 2009) with a beam width of 100. On the last iteration, we extract the alignments imposed by the Viterbi parses as the word alignments outputted by the system.

The rule probability function in the BITG induction algorithm $p$ is defined using fixed probabilities for the structural rules, and a translation table $t$ that is trained using IBM model 1 (Brown *et al.*, 1993) in both directions. To calculate the inside probability of a pair of segments, $P\left(A \stackrel{*}{\Rightarrow} x|G\right)$, we use the algorithm described in Saers *et al.* (2009) for the training.

## 4 Experimental Setup

### 4.1 Data

Our experiments are aimed at showing that injecting a crosslingual semantic objective function into a confidence-weighted ITG induction algorithm into early stage learning of SMT systems can help us reduce the need for extremely large corpora as typically used in SMT training. Although Chinese is not a low resource language, we purposely try to simulate low

Table 1: Translation quality comparing three methods used to train Moses hierarchical PBSMT for Chinese-English MT

| System | BLEU | METEOR | TER | WER | PER | CDER |
|---|---|---|---|---|---|---|
| Giza++ based alignment | 23.02 | 4.14 | 59.95 | 60.52 | 55.58 | 59.14 |
| ITG based alignment | 21.82 | 4.32 | 57.86 | 58.68 | 53.90 | 57.38 |
| Semantic confidence-weighted ITG based alignment | **28.97** | **4.35** | **57.80** | **58.55** | **53.50** | **57.14** |

resource conditions, by using a relatively small corpus (IWSLT07). The training set contains 39,953 sentences. The dev set and test set were the same for all systems in order to keep the experiments comparable.

### 4.2 Baselines

We compare the performance of our proposed confidence-weighted alignment to the conventional ITG alignment and to the traditional GIZA++ baseline with grow-diag-final-and to harmonize both alignment directions. We also perform a grid search over the hyper parameters in our proposed model to find the optimal settings.

We tested the different alignments described above by using the standard Moses toolkit (Koehn *et al.*, 2007), and a 6-gram language model learned with the SRI language model toolkit (Stolcke, 2002) to train our model.

## 5 Results

We compared the performance of the semantically confidence-weighted ITG alignment against the GIZA++ baseline and the conventional BITG alignments. We evaluated our MT output using a broad range of metrics including BLEU (Papineni *et al.*, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006). We note that the alignment based on our proposed algorithm helps to achieve high scores in terms of surface based metrics in comparison to both conventional ITG and GIZA++ alignment. Table 1 shows that our proposed algorithm produces improvements in terms of nearly all metrics, compared to the two conventional alignments. This shows that we should be more focused on incorporating semantic information during the actual learning of the translation model's structure than just tuning against a semantic objective function.

Figure 2 shows examples extracted from our translated data, it compares the translations obtained by the three discussed alignments. We see from the examples that ITG based models can produce semantically more accurate output compared to GIZA++ based alignment. Example 1 shows an interesting example where the confidence-weighted based system learns a more accurate and fluent translation of the input sentence in comparison to both other systems. Example 2 shows an example where learning the right semantic structure can not only produce better adequacy, but also leads to better fluency for low resource languages. The semantic frame based objective function that we used shows that by capturing the right structure while learning the alignment, we can produce better translations even when using a very small data set. This also shows, that semantic based heuristics are needed for more disambiguation, on the other hand, GIZA++ based alignment fails to completely capture any meaning once again.

## 6 Conclusion

In this paper we have introduced a novel crosslingual semantically driven algorithm for inversion transduction grammar induction, where we measure the confidence of the training set based on an XMEANT objective and boost confidence on training examples accordingly. Results suggest that this method of incorporating a semantic frame based objective during early stage of learning a translation model's structure for SMT helps to improve both the fluency and the adequacy of the MT translation, compared to ordinary ITG and to conventional GIZA++ based induction methods.

The performance of our model was tested upon a Moses hierarchical translation baseline. We noted that systems using our early stage semantically based learning approach outperform both conventional GIZA++ and BITG alignment systems in terms of a broad range of metrics including BLEU, METEOR, TER, WER, PER and CDER for Chinese to English translations. We believe this new approach to semantically confidence-weighted training could be conveniently applied to numerous SMT approaches aside from ours.

## 7 Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under LORELEI contract HR0011-15-C-0114, BOLT contracts HR0011-12-C-0014 and HR0011-12-C-0016, and GALE contracts HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the Horizon 2020 grant agreement 645452 (QT21) and FP7 grant agreement 287658; and by the

**Example 1**

| | |
|---|---|
| ***Input:*** | 我 在 这 家 公司 工作 九 年 了 所 以 今年 有 四 个 星期 的 带薪 休假 。 |
| ***Gloss:*** | I in this company works nine years so this year have four week paid vacation. |
| ***Ref:*** | I have been with our Company for nine years and I am entitled to four weeks of paid leave this year. |
| ***GIZA++:*** | I work at this company nine years have four weeks VACATION this year. |
| ***ITG:*** | I work at this company nine years by four weeks paid vacation this year. |
| ***Confidence-weighted:*** | I work in this nine years , so let 's have four weeks paid vacation. |

**Example 2**

| | |
|---|---|
| ***Input:*** | 食堂 在 哪里 ？ |
| ***Gloss:*** | canteen at where? |
| ***Ref:*** | where 's the dining room ? |
| ***GIZA++:*** | refectory then where ? |
| ***ITG:*** | the refectory where ? |
| ***Confidence-weighted:*** | where is the refectory ? |

Figure 2: Examples comparing the output of the three discussed alignments

Hong Kong Research Grants Council (RGC) research grants GRF16210714, GRF16214315, GRF620811 and GRF621008. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.

## References

Karteek Addanki, Chi-kiu Lo, Markus Saers, and Dekai Wu. LTG vs. ITG coverage of cross-lingual verb frame alternations. In *16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, Trento, Italy, May 2012.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005.

Meriem Beloucif, Chi kiu Lo, and Dekai Wu. Improving meant based semantically tuned smt. In *11 th International Workshop on spoken Language Translation (IWSLT 2014), 34-41 Lake Tahoe, California*, 2014.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederik Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

Adrià De Gispert, Deepa Gupta, Maja Popovic, Patrik Lambert, Jose B.Marino, Marcello Federico, Hermann Ney, and Rafael Banchs. Improving statistical word alignment with morpho-syntactic transformations. In *Advances in Natural Language Processing*, pages 368–379, 2006.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *The second international conference on Human Language Technology Research (HLT '02)*, San Diego, California, 2002.

Chris Dyer, Jonathan Clark, Alon Lavie, and Noah A.Smith. Unsupervised word alignment with arbitrary features. In *49th Annual Meeting of the Association for Computational Linguistics*, 2011.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Interactive Poster and Demonstration Sessions of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic, June 2007.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

Chi-kiu Lo and Dekai Wu. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, 2011.

Chi-kiu Lo and Dekai Wu. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

Chi-kiu Lo and Dekai Wu. Can informal genres be better translated by tuning on automatic semantic metrics? In

*14th Machine Translation Summit (MT Summit XIV)*, 2013.

Chi-kiu Lo and Dekai Wu. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based mt evaluation metric. In *8th Workshop on Statistical Machine Translation (WMT 2013)*, 2013.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully automatic semantic MT evaluation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu. Improving machine translation by training against an automatic semantic frame based evaluation metric. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

Chi-kiu Lo, Meriem Beloucif, and Dekai Wu. Improving machine translation into Chinese by tuning against Chinese MEANT. In *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. XMEANT: Better semantic MT evaluation without reference translations. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 2014.

Jeff Ma, Spyros Matsoukas, and Richard Schwartz. Improving low-resource statistical machine translation with a novel semantic word clustering algorithm. In *Proceedings of the MT Summit XIII*, 2011.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In *Eighth Workshop on Statistical Machine Translation (WMT 2013)*, Sofia, Bulgaria, August 2013.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A evaluation tool for machine translation: Fast evaluation for MT research. In *The Second International Conference on Language Resources and Evaluation (LREC 2000)*, 2000.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *The 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 440–447, Hong Kong, October 2000.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania, July 2002.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow semantic parsing using support vector machines. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 2004.

Markus Saers and Dekai Wu. Improving phrase-based translation via word alignments from stochastic inversion transduction grammars. In *Third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, pages 28–36, Boulder, Colorado, June 2009.

Markus Saers, Joakim Nivre, and Dekai Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *11th International Conference on Parsing Technologies (IWPT'09)*, pages 29–32, Paris, France, October 2009.

Markus Saers, Joakim Nivre, and Dekai Wu. Word alignment with stochastic bracketing linear inversion transduction grammar. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 341–344, Los Angeles, California, June 2010.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, August 2006.

Theerawat Songyot and David Chiang. Improving word alignment using word similarity. In *52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, pages 901–904, Denver, Colorado, September 2002.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *The 16th International Conference on Computational linguistics (COLING-96)*, volume 2, pages 836–841, 1996.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997.

Hao Zhang and Daniel Gildea. Stochastic lexicalized inversion transduction grammar for alignment. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 475–482, Ann Arbor, Michigan, June 2005.