The CMU-UKA Statistical Machine Translation Systems for IWSLT 2007

Ian Lane[*]*, Andreas Zollmann*[*]*, Thuy Linh Nguyen*[*]*, Nguyen Bach*[*]*,*
Ashish Venugopal[*]*, Stephan Vogel*[*]*, Kay Rottmann*[†]*, Ying Zhang*[*] *and Alex Waibel*[†*]

InterACT Research Laboratories:
[*]Carnegie Mellon University, Pittsburgh, USA [†]University of Karlsruhe, Karlsruhe, Germany

Abstract

This paper describes the CMU-UKA statistical machine translation systems submitted to the IWSLT 2007 evaluation campaign. Systems were submitted for three language-pairs: Japanese→English, Chinese→English and Arabic→English. All systems were based on a common phrase-based SMT (statistical machine translation) framework but for each language-pair a specific research problem was tackled. For Japanese→English we focused on two problems: first, punctuation recovery, and second, how to incorporate *topic*-knowledge into the translation framework. Our Chinese→English submission focused on syntax-augmented SMT and for the Arabic→English task we focused on incorporating morphological-decomposition into the SMT framework. This research strategy enabled us to evaluate a wide variety of approaches which proved effective for the language pairs they were evaluated on.

1. Introduction

For the IWSLT 2007 evaluation campaign we focused on applying systems developed based on current research topics within our lab. This includes our work on syntax-augmented SMT (statistical machine translation), morphological-decomposition for rich morphology languages, such as Arabic, and approaches to better couple ASR (automatic speech recognition) and MT (machine translation) for spoken language translation. Three systems were submitted for evaluation: Japanese→English, Chinese→English and Arabic→English, and all built upon our 2006 submissions for these tasks. This year we focused on applying a number of distinct approaches for each language-pair. This enabled us to cover many research topics within spoken language translation.

In Section 3, we describe our J→E submission system and compare various approaches to recover punctuation (Section 3.3). In Section 3.4 we investigate methods to incorporate *topic*-knowledge into the translation framework via N-best list re-scoring. Our syntax-augmented SMT framework is detailed in Section 4. Sections 4.5 and 4.6 describe its application to the C→E task. Our A→E translation system which incorporates morphological decomposition is introduced in Section 5.

2. The CMU-UKA Phrase-based SMT System

Our J→E and A→E systems built upon the STTK (Statistical Translation Toolkit) framework used for our IWSLT 2006 submissions [1]. STTK implements phrase-based statistical machine translation using a log-linear model [2] in which a foreign language sentence $f_1^J = f_1, f_2, \ldots, f_J$ is translated into another language $e_1^I = e_1, e_2, \ldots, e_I$ by searching for the hypothesis $\hat{e}_1^I$ with maximum likelihood, given:

$$\begin{aligned}\hat{e}_1^I &= \arg\max_{e_1^I} P(e_1^I|f_1^J) \\ &= \arg\max_{e_1^I} \frac{exp\left(\sum_{m=1}^M \lambda_m h_m\left(e_1^I, f_1^J\right)\right)}{\sum_{e'^{I'}_1} exp\left(\sum_{m=1}^M \lambda_m h_m\left(e'^{I'}_1, f_1^J\right)\right)}\end{aligned}$$

In this framework, the posterior probability $P(e_1^I|f_1^J)$ is directly maximized using a log-linear combination of feature functions $h_m\left(e_1^I, f_1^J\right)$ (during decoding the denominator is dropped since it depends only on f_1^J). Feature functions applied during translation include: language models, translation models, and sentence length models. The scaling factors $(\lambda_1, \lambda_2, \ldots, \lambda_M)$ applied during search are optimized via MERT (minimum error rate training) [3] for a specific translation metric such as BLEU [4]. Search is performed using our STTK beam-search-decoder [5] which allows restricted word re-ordering during translation.

3. Topic-Aware Japanese-to-English SLT

For the Japanese-to-English submission we focused on two research areas. First, we compared various methods to recover intra-utterance sentence boundaries and secondary punctuation (commas); and second, we investigated approaches to incorporate *topic*-knowledge into the SMT framework. These works are described in Sections 3.3 and 3.4, respectively. By incorporating publicly available corpora from related domains and applying the proposed techniques the translation accuracy of our system improved significantly. On the 2006 IWSLT J→E evaluation task our 2007 system obtained obtaining 0.2546 BLEU, compared to 0.2030 for our 2006 submission (open-data track, correct recognition result task).

Table 1: *Dev. and Held-out Eval. Sets for J→E Task*

Development Set: IWSLT JE devset4	
Description: 2005 IWSLT J→E evaluation set	
No Utterances: 489 (6491 Source Tokens)	
Held-out Evaluation Set: IWSLT JE devset5	
Description: 2006 IWSLT J→E evaluation set	
No Utterances: 500 (7113 Source Tokens)	

3.1. Training Corpora Selection

To improve system performance we investigated using publicly available corpora from related domains for this task. Five publicly available, J-E corpora (indicated in Table 2) were evaluated for relevance to the IWSLT task-domain. Corpora were preprocessed by removing unnecessary punctuation (only commas, full-stops and question-marks were retained) and numerals were converted to their spoken form on both the source (Japanese) and target (English) sides. Japanese word-segmentation was performed using the Conditional-Random-Field based morphological analyzer "*mecab*" [8], and sentence-pairs with outlying source-target token ratios were removed.

Various metrics were evaluated for each corpus to determine their relevance to the IWSLT task-domain. These included: average sentence-length, OOV (out-of-vocabulary) rate, and target-side perplexity. Target-side perplexity is the perplexity of an n-gram LM trained on that corpora and evaluated on the English references of the development-set. The resulting metrics for each corpus (evaluated using the development-set described in Table 1) are shown in Table 2. The IWSLT corpora and "*Tanaka Corpus*" obtained low OOV-rates and low target-side perplexity, indicating the relevance of these corpora for this task. The average sentence length of the "*JENAAD*" and "*Reuters*" corpora was three times longer than that of the IWSLT corpora and target-side-perplexity was significantly higher. These two corpora were determined not to be relevant [1]. Combining the relevant corpora (Table 2, "*Combined*") significantly reduced the OOV-rate (from 1.6% to 0.7%) and target-side perplexity (from 71.5 to 65.3) compared to using the provided IWSLT data alone. This combined corpus consisting of 2.4M source-tokens was used to develop the J→E system.

3.2. Baseline System

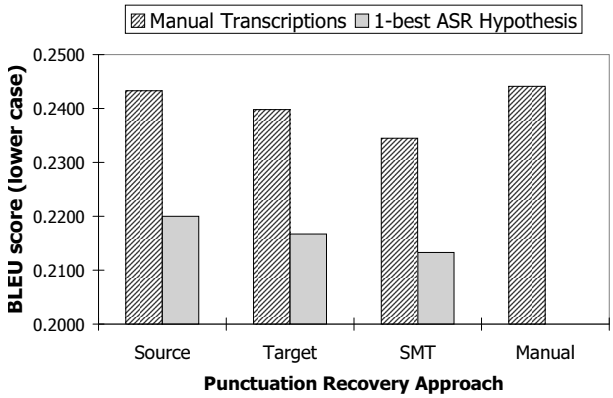
A phrase-based J→E SMT system was developed using the same framework as our 2006 system [1]. Phrase extraction was performed using the *PESA* (Phrase Pair Extraction as Sentence Splitting) method proposed in [9]. SMT decoding was performed using our STTK decoder described in Section 2. Two target language models were applied during decoding: a 6-gram suffix-array language model trained on the combined corpora and a 4-gram LM, with Kneser-Ney smoothing, obtained by interpolating LMs trained on the three individual corpora to minimized perplexity on the development-set. A re-ordering window of 6 was applied during decoding. Scaling factors for feature functions were optimized via MERT on the development-set. Translation was performed using a lower-case target (English) and case-recovery was performed using noisy-channel-model.

The baseline system obtained a case-sensitive BLEU score of 0.2172 on the held-out evaluation set.

[1] Incorporating the two "*non-relevant corpora*" slightly degraded translation-quality (BLEU) on the held-out evaluation set (from 0.2286 to 0.2304 (case sensitive)). However, other measures of translation-quality, specifically, BLEU-precision and TER improved when this data was included, indicating that the above metrics may be of limited use for corpora selection.

Table 3: *Accuracy of Source-side Punctuation Recovery on Held-Out Evaluation Set*

	Precision	Recall	F-score
Manual transcripts			
Sentence Boundary	97.8%	96.8%	97.3%
Secondary Punctuation	82.1%	44.2%	57.5%
1-best ASR Hypothesis: *Character Error Rate=10.4%*			
Sentence Boundary	96.4%	95.9%	96.2%
Secondary Punctuation	71.8%	43.6%	54.3%
Lower Bound: *assume sentence boundary at end of utterance*			
Sentence Boundary	100%	63.9%	77.9%



Source: Punctuation recovery applied to input source
Target: Punctuation recovery applied after translation (secondary-punctuation ignored during translation)
SMT: Punctuation recovered during translation
Manual: Input source manually annotated with punctuation

Figure 1: *Comparison of punctuation recovery schemes*

3.3. Punctuation Recovery for J→E SLT

The handling of punctuation in SLT (spoken language translation) is important not only to improve the readability of the translation output, but also for its role during word-alignment and phrase-extraction. One traditional approach to recover punctuation in SLT is to disregard it during translation and then re-generate it on the target side via a HELM (hidden-

Table 2: *Training Corpora used for the UKA/CMU Japanese-to-English Translation System*

	Corpora Name	No. Sent. Pairs	No. Tokens Source	No. Tokens Target	Avg. Sent. Length Source	Avg. Sent. Length Target	OOV Rate (Source)	Target-side ppl (Vocab Size=70k)
1	**IWSLT provided**	39,953	429,010	381,776	10.74	9.56	1.6%	71.5
2	**IWSLT devsets 1-3**	24,192	212,256	204,387	8.77	8.45	10.4%	128.5
3	**Tanaka Corpora [6]**	155,340	1,706,683	1,367,354	11.52	9.23	1.8%	119.1
	JENAAD [7]	179,299	6,233,303	5,477,149	24.64	30.44	3.0%	269.1
	Reuters [7]	66,284	2,603,865	1,998,587	37.13	28.50	12.1%	415.0
	Combined (1, 2, 3)	219,485	2,356,949	1,953,517	10.92	9.06	0.7%	65.3

event language model). This technique was applied in [10]. However, there are three drawbacks to this approach: first, this approach is limited by the translation accuracy of the system; second, as punctuation landmarks are discarded, the accuracy of word-alignment will be degraded during phrase extraction; and third, it is non-trivial to incorporate acoustic cues when estimating target-side punctuation. To improve the accuracy of our translation system we compared three approaches to automatically recover punctuation during translation: HELM-based punctuation recovery on the source or target-side, and punctuation recovery via SMT.

First, the effectiveness of punctuation recovery on the input source was evaluated. A HELM was trained on the source-side of the training corpora to estimate both sentence-boundaries and secondary punctuation (in this work only commas are considered). This model was then applied to the manual transcripts and 1-best ASR output of the held-out evaluation set. Accuracy was evaluated by comparing the output hypothesis to manually annotated reference. Precision, recall and F-score are shown in Table 3. Punctuation recovery using a HELM obtained F-scores of 97.3% and 96.2% for sentence boundaries, when applied to the manual transcripts and 1-best hypotheses, respectively. For secondary punctuation (commas) F-scores of 57.5%, and 54.3% were obtained.

Source-side recovery of sentence-boundaries via a HELM obtained high accuracy and performance was not significantly degraded by speech recognition errors. Thus, this approach was applied in the remaining experiments. For secondary punctuation, however, although precision was high, recall was below 50%. This could be due to inconsistent annotation between the training corpora and evaluation set.

Next, to improve the handling of secondary punctuation within SMT the effectiveness of three approaches were compared: HELM-based punctuation recovery on the source ("*Source*") or target ("*Target*") sides, and estimation of target-side punctuation via SMT "*SMT*", which involves translating from a source with no secondary punctuation to a target with full punctuation. For comparison the performance for manually annotated punctuation ("*Manual*") was also evaluated. The performance of these four systems are shown in Figure 1 when applied to the manual transcripts and 1-best ASR hypotheses.

For the J→E IWSLT task, source-side punctuation-recovery obtained the highest translation performance for both the manual transcription and ASR cases. Furthermore, this approach obtained performance close to that obtained by manually annotation. Source-side punctuation-recovery seems to be effective for this task due to the limited domain. Also, the high ASR-accuracy of this task limits degradation due to recognition errors. In future work, we intend to extent this approach to incorporate prosodic features, and to improve robustness by considering source-side punctuation in probabilistic manner within SMT decoding.

3.4. Incorporating Topic-Knowledge into SMT

Table 4: *Effectiveness of N-best list rescoring with different feature sets on Held-out Evaluation Set (BLEU lowercase)*

Baseline	TDLM	TC	TDLM + TC
0.2432	0.2662	0.2709	0.2678

TDLM: *Topic*-dependent language model
TC: *Topic*-confidence scores

To incorporate *topic*-knowledge into the SMT framework we investigated two sets of *topic*-features: topic-dependent language model score and topic confidence scores. These features were calculated over an entire translation hypothesis and thus could not be easily incorporated into the beam-search-decoder. Instead, a large N-best list was generated in the first pass, and this was re-scored using feature-scores generated during the 1st-pass decoding and the additional *topic*-features described below. The scaling factors for the combined feature-set were optimized via MERT on the development set.

The first feature we investigated was a topic-dependent LM score "*TDLM*". For each utterance a single *topic*-class was selected based on the topic classification result of the 1-best translation result. The log probability of the relevant topic-dependent LM was then calculated and used as an additional feature during N-best list re-scoring. The assumption is that a topic-dependent LM will better discriminate between acceptable and bad translations for a specific topic-class than a background language model trained over the entire corpus.

The second set of features we investigated were topic-

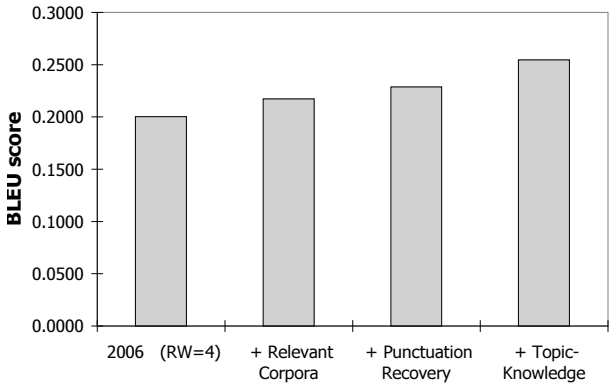


Figure 2: *Improvements of the J→E system on the IWSLT2006 evaluation set*

confidence scores "*TC*". For each translation hypothesis topic-confidence scores were calculated for all topic classes. These scores were then added as additional features during re-scoring. Topic confidence scores were generated using SVM-based topic classifiers. Rather than only considering word identity features during classification, 2-gram, 3-gram features were also incorporated. It was observed that fluent translation hypotheses obtained high topic-confidence for a single *topic*-class, where as, poor translations, especially those that were semantically incoherent, typically had low confidence across all classifiers. In this manner the topic-classification scores provide a measure of topic-consistency.

As topic labels did not exist for the training corpora, we generated pseudo *topic*-classes by applying hierarchical clustering to the target-side of the training data. Clustering was performed to minimize overall perplexity. For each *topic*-cluster a 4-gram LM and SVM-based topic classifier were trained. In the evaluation system eight *topic*-classes were used. The effectiveness of N-best list rescoring incorporating the two sets of features described above are shown in Table 4.

On the held-out evaluation set (devset5), the highest translation score was obtained by incorporating topic-confidence features during re-scoring ("*TC*"). Incorporating the topic-dependent LM feature improved performance compared to the baseline system, however, this feature tended to over-generate words in the translation hypothesis, thus degrading performance when combined with the "*TC*" features.

3.5. The CMU-UKA J→E Submission System

For the CMU-UKA J→E submission system, first, source-side punctuation recovery was applied using the approach described in Section 3.3. Next, 1000-best translation hypotheses were generated for each input utterance using the phrase-based SMT system described in Section 3.2. Finally, N-best list rescoring was performed. In this step, topic-consistency scores were computed independently for each translation hypothesis, and these scores were incorporated into the log-linear translation model. The hypothesis with maximum likelihood after re-scoring was output as the final translation result. For the spoken language translation task, MERT was applied to the 1-best ASR hypotheses from the development set rather than to the manual transcriptions as was the case for the correct recognition result task.

Our 2007 system significantly improved translation quality compared to our 2006 submission. Figure 2 shows the improvements gained for each approach.

4. Syntax Augmented SMT for Chinese-to-English Translation

There is currently intense interest in the application of hierarchical and syntax driven models for statistical machine translation. These models seek to address the problem of generating fluent, well structured target language output under the premise that human language is essentially hierarchical in its generation. Hierarchical approaches gain their representational power by allowing transformation rules to condition on larger fragments of target language tree structure. The application of hierarchically structured models to statistical machine translation requires the development of techniques to induce and estimate transformation rules from parallel data (grammar induction), and efficient algorithms to apply these rules to translate source language text (decoding).

In recent work [11], we presented the first results that leverage target language syntactic structure to achieve higher performance than comparable phrase based translation. [12] presents results that show the impact of hierarchical structure alone, and [13] achieves significant improvements using tree-to-string transformations.

For the Chinese-to-English task, we used the latest version of the Syntax-Augmented Machine Translation (SAMT) system first described in [11]. The system is available open-source under the GNU General Public License at:
`www.cs.cmu.edu/~zollmann/samt`

4.1. Synchronous Grammars for SMT

Probabilistic synchronous context-free grammars (PSCFGs) are defined by a source terminal set (source vocabulary) $\mathcal{T}_S$, a target terminal set (target vocabulary) $\mathcal{T}_T$, a shared nonterminal set $\mathcal{N}$ and induce rules of the form

$$X \rightarrow \langle \gamma, \alpha, \sim, w \rangle$$

where

- $X \in \mathcal{N}$ is a nonterminal,
- $\gamma \in (\mathcal{N} \cup \mathcal{T}_S)^*$ is a sequence of nonterminals and source terminals,
- $\alpha \in (\mathcal{N} \cup \mathcal{T}_T)^*$ is a sequence of nonterminals and target terminals,
- the count $\#\mathrm{NT}(\gamma)$ of nonterminal tokens in γ is equal to the count $\#\mathrm{NT}(\alpha)$ of nonterminal tokens in α,

- $\sim$: $\{1, \ldots, \#\mathrm{NT}(\gamma)\} \rightarrow \{1, \ldots, \#\mathrm{NT}(\alpha)\}$ is a one-to-one mapping from nonterminal tokens in γ to nonterminal tokens in α, and
- $w \in [0, \infty)$ is a nonnegative real-valued weight assigned to the rule.

In our notation, we will assume $\sim$ to be implicitly defined by indexing the NT occurrences in γ from left to right starting with 1, and by indexing the NT occurrences in α by the indices of their corresponding counterparts in γ. Syntax-oriented PSCFG approaches often ignore source structure, instead focusing on generating syntactically well-formed target derivations. [12] use a single nonterminal category, [14] use syntactic constituents for the PSCFG nonterminal set, and [11] take advantage of CCG [15] inspired "slash" and "plus" categories.

4.2. Grammar Induction

The SAMT model generates a PSCFG given parallel sentence pairs $\langle f, e \rangle$, a parse tree π for each e, the maximum *a posteriori* word alignment a over $\langle f, e \rangle$, and a set of phrase pairs $Phrases(a)$ identified by any alignment-driven phrase induction technique such as e.g. [16].

Each phrase in $Phrases(a)$ is first annotated with a syntactic category to produce initial **rules**, where γ is set to the source side of the phrase, α is set to the target side of the phrase, and X is assigned based on the corresponding target side span in π. If the target span of the phrase does not match a constituent in π, heuristics are used to assign categories that correspond to partial rewriting of the tree. These heuristics first consider concatenation operations, forming categories like "NP+VP", and then resort to CCG style "slash" categories like "NP/NN." Preference for the concatenation operations over the slash categories is based on the assumption that categories closer to the leaves of the tree are more accurate and more strongly tied to the words than categories higher up the tree.

To illustrate this annotation process, we consider the following French-English sentence pair and selected phrase pairs obtained by phrase induction on an automatically produced alignment a:

f	=	il ne va pas
e	=	he does not go
il	:	he
va	:	go
ne va pas	:	not go
il ne va pas	:	he does not go

The alignment a with the associated target side parse tree is shown in Fig. 3 in the alignment visualization style defined by [14]. Matching the target span of each phrase with the parse π, we generate the following initial rules.

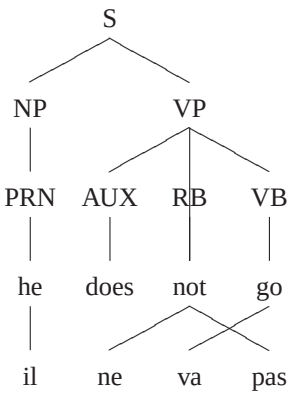


Figure 3: Alignment graph (word alignment and target parse tree) for a French-English sentence pair.

PRP	$\rightarrow$	il, he
VB	$\rightarrow$	va, go
RB+VB	$\rightarrow$	ne va pas, not go
S	$\rightarrow$	il ne va pas, he does not go

Note that the third rule illustrates the use of concatenation categories to identify syntactic categories. These initial rules form the lexical basis for generalized rules that include labeled syntactic categories in γ and α. Following the Data-Oriented Parsing [17] inspired rule generalization technique proposed by [12], one can now generalize each **identified** rule (initial or already partially generalized)

$$N \rightarrow f_1 \ldots f_m / e_1 \ldots e_n$$

for which there is an **initial** rule

$$M \rightarrow f_i \ldots f_u / e_j \ldots e_v$$

where $1 \leq i < u \leq m$ and $1 \leq j < v \leq n$, to obtain a new rule

$$N \rightarrow f_1 \ldots f_{i-1} M_k f_{u+1} \ldots f_m / e_1 \ldots e_{j-1} M_k e_{v+1} \ldots e_n$$

where k is an index for the nonterminal M that indicates the one-to-one correspondence between the new M tokens on the two sides (it is not in the space of word indices like i, j, u, v, m, n). The recursive form of this generalization operation allows the generation of rules with multiple nonterminal symbols. Note that since we only generalize over initial rules, this operation has polynomial runtime as a function of $|Phrases(a)|$.

The initial rules listed above can be generalized to additionally extract the following rules from f, e.

$$
\begin{aligned}
\text{S} &\rightarrow \text{PRP}_1 \text{ ne va pas }, \text{ PRP}_1 \text{ does not go} \\
\text{S} &\rightarrow \text{il ne VB}_1 \text{ pas }, \text{ he does not VB}_1 \\
\text{S} &\rightarrow \text{il RB+VB}_1, \text{ he does RB+VB}_1 \\
\text{S} &\rightarrow \text{PRP}_1 \text{ RB+VB}_2, \text{ PRP}_1 \text{ does RB+VB}_2 \\
\text{RB+VB} &\rightarrow \text{ne VB}_1 \text{ pas }, \text{ not VB}_1
\end{aligned}
$$

4.3. Decoding

Given a source sentence f, the translation task under a PSCFG grammar can be expressed analogously to monolingual parsing with a CFG. We find the most likely derivation D of the input source sentence while reading off the English translation from this derivation:

$$\hat{e} = \text{tgt}\left(\underset{D:\text{src}(D)=f}{\arg\max}\ p(D)\right) \quad (1)$$

where $\text{tgt}(D)$ refers to the target terminal symbols generated by the derivation D and $\text{src}(D)$ refers to the source terminal symbols spanned by D.

Our distribution p over derivations is defined by a log-linear model. The probability of a derivation D is defined in terms of the rules r that are used in D:

$$p(D) = \frac{p_{LM}(\text{tgt}(D))^{\lambda_{\text{LM}}} \times \prod_{r \in D} \prod_i \phi_i(r)^{\lambda_i}}{Z(\lambda)} \quad (2)$$

where ϕ_i refers to features defined on each rule, p_{LM} is a n-gram LM probability applied to the target terminal symbols generated by the derivation D, and $Z(\lambda)$ is a normalization constant chosen such that the probabilities sum up to one. The computational challenges of this search task (compounded by the integration of the language model) are addressed elsewhere [18, 19]. All feature weights λ_i are trained in concert with the language model weight via minimum-error training [3]. Here, we focus on the estimation of the feature *values* ϕ during the grammar induction process. The feature values are statistics estimated from rule counts.

4.4. Feature Value Statistics

The features ϕ represent multiple criteria by which the decoding process can judge the quality of each rule and, by extension, each derivation. We include both real-valued and boolean-valued features for each rule. The following probabilistic quantities are estimated and used as feature values:

- $\hat{p}(r|\,\text{lhs}(X))$: Probability of a rule given its l.h.s category
- $\hat{p}(r|\,\text{src}(r))$: Probability of a rule given its source side
- $\hat{p}(r|\,\text{tgt}(r))$: Probability of a rule given its target side
- $\hat{p}(\text{ul}(\text{src}(r)), \text{ul}(\text{tgt}(r))|\,\text{ul}(\text{src}(r)))$: Probability of the unlabeled source and target side of the rule given its unlabeled source side.
- $\hat{p}(\text{ul}(\text{src}(r)), \text{ul}(\text{tgt}(r))|\,\text{ul}(\text{tgt}(r)))$: Probability of the unlabeled source and target side of the rule given its unlabeled target side.

where lhs returns the left-hand-side of a rule, src returns the source side γ, and tgt returns the target side α of a rule r. The function ul removes all syntactic labels from its arguments, but retains ordering notation. For example, $\text{ul}(\text{NP+AUX}_1\text{does not go}) = \square_1$ does not go. The last two features are extensions to the feature set suggested by [11]. They represent the same kind of relative frequency estimates commonly used in phrase based systems. The ul function allows us to calculate these estimates for rules with nonterminals as well.

To estimate these probabilistic features, we use maximum likelihood estimates based on counts of the rules extracted from the training data. For example, $\hat{p}(r|lhs(r))$ is estimated by computing $\#(r)/\#(\text{lhs}(r))$, aggregating counts from all extracted rules.

As in phrase-based translation model estimation, ϕ also contains two lexical weights $\hat{p}_w(\text{lex}(\text{src}(r))|\,\text{lex}(\text{tgt}(r)))$ and $\hat{p}_w(\text{lex}(\text{tgt}(r))|\,\text{lex}(\text{src}(r)))$ [20] that are based on the lexical symbols of γ, α. These weights are estimated based on an pair of statistical lexicons that represent $\hat{p}(s|t), \hat{p}(t|s)$, where s and t are single words in the source and target vocabulary. These word-level translation models are typically estimated by maximum likelihood, considering the word-to-word links from "single-best" alignments as evidence.

We also store several boolean and count features in ϕ: the rule is purely lexical in α and γ; the rule is purely *non-lexical* in α and γ; the number of target words in the rule.

4.5. Training Corpora

We used the provided 40K sentence BTEC corpus for training. As successfully employed by some teams in last year's competition, we added the development sets (except devset4, which we used for MER training) to the training corpus as well. Since each devset source sentence f corresponds to several references $e_1, \ldots, e_n$, we added one sentence pair (f, e_i) for each reference e_i. The resulting training corpus comprised of 67645 sentence pairs total.

4.6. The CMU-UKA C→E Submission System

The submitted system was MER-tuned to devset4 towards the NIST-BLEU metric for 5 iterations, yielding a final score of 32.5%. The official score of the system during the evaluation was 37.44%. We assume that this was according to the IBM-BLEU metric.

5. Morphological-Decomposition for Robust Arabic-to-English SMT

Statistical machine translation relies on a word alignment model, between source and target language, to extract and score phrase translations. In current word alignment methods

Training		
IWSLT-BTEC	19847 sent.	157,795 Arabic words 189,861 English words
Development Data		
IWSLT 04	500 sent.	16 references
IWSLT 05	506 sent.	16 references

Table 5: Arabic - English Data

[2, 21] the one-to-one mapping between tokens in the source and target language is critical. However, for very diverse language pairs, i.e. translating between a rich morphology language such as Arabic and a poor morphology language such as English, a significant mismatch is present. For this language-pair, prefixes and suffixes of an Arabic word often correspond to separate English words. When translating from Arabic to English, a preprocessing step on Arabic is necessary to maintain consistency between two languages.

In our A→E submission system we applied full morphological decomposition to the training corpora using a state-of-the-art Arabic morphological analyzer [22]. Morphological decomposition replaces each word in the training corpora with a sequence of its component morphemes *prefix-*, *stem*, *-suffix*. This approach also improves the coverage of the system, enabling it to translate words that do not occur in the training data, by performing translation at the sub-word, morpheme level.

The prefix of an Arabic word can be a combination of conjunction (*wa* - and), article (*Al* - the), and preposition (*li* - to/for). Its' suffix can be a pronoun (*hm* - their/them), case marker (*u, i, a*) gender (*f* - female singular), number, or voice, etc. Even though many morphemes have an equivalent English translation, some specific morphemes like gender, number, and case markers are redundant and can be discarded when translating to English. Previous work [23], used knowledge of the Arabic language to explicitly remove inflectional features from Arabic text before translation.

In this work, we attempt to discard non-informative morphemes using a data driven approach. First, Arabic full morphology analysis and English text are passed to Giza++ word alignment toolkit [24]. Figure 4 shows the fertility distributions of Arabic morphemes from this alignment for the training corpora defined below. Morphemes that are not aligned to any English word are indicated by high zero-fertility probability. These morphemes must be discarded to reduce NULL alignments and balance sentence length between the Arabic and English data. Morphemes for which the zero-fertility are greater than a threshold θ_{th} are discarded from the Arabic text. The threshold θ_{th} is selected to maximize translation quality (BLEU) on a development set.

In the experimental evaluation, the "*IWSLT A→E provided data*" was used for system training. This corpus consists of about 20K sentence pairs. The IWSLT04 and IWSLT05 A→E evaluation sets were used as development sets. Details of the training and development data are described in Table 5. From $157,795$ original Arabic words, we obtained $294,678$ morphemes. This compares to $189,861$ English words. Phrase extraction was performed using the Pharaoh training scripts [20], our STTK decoder, described in Section 2, was applied for decoding.

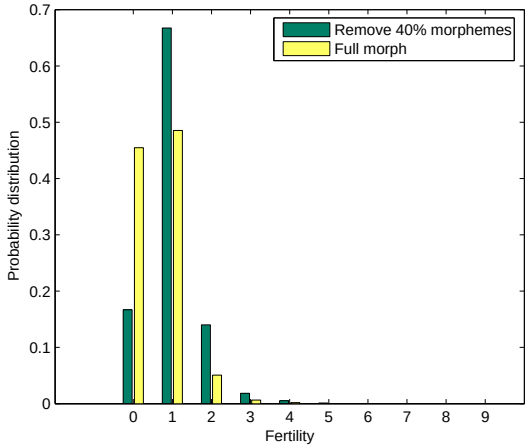


Figure 4: Fertility distributions of full morphology analysis Arabic and remove 40% word tokens morphology Arabic

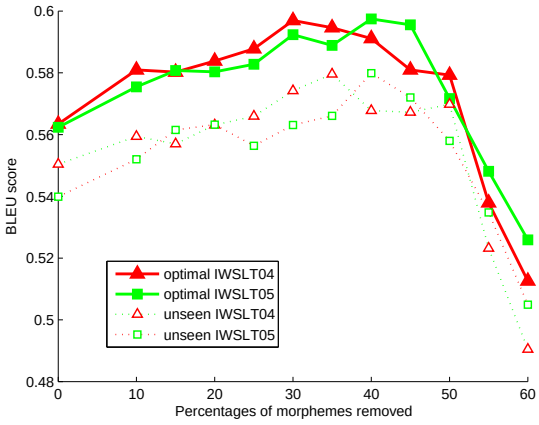


Figure 5: Translation result as a function morpheme token rates remove from Arabic text

In our experiments, IWSLT04 and IWSLT05 are alternatively used one as a development set and the other as an unseen test set. Figure 5 shows the translation quality (BLEU) when different thresholds of θ_{th} are applied to discard morphemes. The threshold also relates to the percentage of morphemes discarded. In Figure 5, the solid lines indicate the translation accuracy of the development-set and the dotted lines indicate the performance on the unseen test set. When IWSLT04 was applied as the development set, the highest BLEU score was obtained when 30% of morphemes were discarded. For the IWSLT05 case discarding 40% of morphemes obtained the highest score.

	Test Set	Original Arabic	With Morphology
Dev. IWSLT04	IWSLT04	0.5758	0.5970 (+2.1)
(*remove 30%*)	IWSLT05	0.5573	0.5631 (+0.8)
Dev. IWSLT05	IWSLT04	0.5583	0.5742 (+1.6)
(*remove 40%*)	IWSLT05	0.5756	0.5974 (+2.2)

Table 6: IWSLT04 and IWSLT05 results

Table 6 compares the translation quality (BLEU) of the baseline system, trained using the original Arabic word-tokens only, and the proposed approach, incorporating morphological decomposition. On the held-out evaluation sets, an improvement in BLEU of 0.8 and 1.6 points were obtained for the IWSLT05 and IWSLT04 sets, respectively.

5.1. The CMU-UKA A→E Submission System

The above system was applied to the IWSLT 2007 A→E spoken language translation and clean text tasks. Punctuation normalization and morphological decomposition was applied to the clean Arabic text. For the clean text task the IWSLT05 set was used for parameter tuning and 40% of morphemes were discarded. By incorporating morphological decomposition into the SMT framework our A→E submission system obtained a BLEU-score of 0.4463 on the IWSLT 2007 "*clean text*" evaluation set. For the spoken language translation task, the IWSLT06 set was used for parameter tuning and on this task, our A→E SMT system obtained a BLEU-score of 0.3756.

6. Conclusion

In this IWSLT evaluation, we investigated several new approaches: source-side punctuation recovery, N-Best list rescoring incorporating *topic*-confidence scores, morphological decomposition for A→E translation and Syntax-Augmented SMT. These approaches proved effective for the language-pairs they were evaluated on and we expect further improvement by combining these techniques.

Acknowledgments: The work reported here was partly funded by the National Science Foundation (NSF) under the project STR-DUST (Grant number IIS-0325905). The work in Morphologically Aware Arabic-to-English Translation was in collaboration with Noah Smith, assistant professor at Language Technologies Institute, Carnegie Mellon University.

7. References

[1] M. Eck, I. Lane, N. Bach, S. Hewavitharana, M. Kolss, B. Zhao, A. S. Hildebrand, S. Vogel, and A. Waibel, "The UKA/CMU statistical machine translation system for IWSLT 2006," in *Procedings of IWSLT*, 2005, pp. 130–137.

[2] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[3] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7 2003.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation." in *Proc. of ACL*, 2002.

[5] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venogupal, B. Zhao, and A. Waibel, "The CMU statistical translation system," in *Proceedings of MT Summit IX*, New Orleans, LA, September 2003.

[6] T. Yasuhito, "Compilation of a multilingual parallel corpus," in *Proceedings of PACLING*, 2001.

[7] M. Utiyama and H. Isahara, "Reliable measures for aligning japanese-english news articles and sentences," in *Proceedings of ACL*, 2003, pp. 72–79.

[8] T. Kudou, "Mecab: Yet another part-of-speech and morphological analyzer," in *http://mecab.sourceforge.net*, 2007.

[9] S. Vogel, "PESA: phrase pair extraction as sentence splitting," in *Proceedings of MT Summit X*, Phuket, Thailand, September 2005.

[10] R. Zhang, H. Yamamota, M. Paul, H. Okuma, K. Yasuda, Y. Lepage, E. Denoual, D. Mochihashi, A. Finch, and E. Sumita, "The NiCT-ATR statistical machine translation system for the iwslt 2006 evaluation," in *Proc. IWSLT*, 2006, pp. 83–90.

[11] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing," in *Proc. of the Workshop on Statistical Machine Translation, HLT/NAACL*, June 2006.

[12] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proc. of ACL*, 2005.

[13] D. Marcu, W. Wang, A. Echihabi, and K. Knight, "SPMT: Statistical Machine Translation with Syntactified Target Language Phrases," in *Proc. of EMNLP*, 2006.

[14] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "What's in a translation rule?" in *Proc. of HLT/NAACL*, 2004.

[15] M. Steedman, "Alternative quantifier scope in CCG," in *Proc. of ACL*, 1999.

[16] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Comput. Linguist.*, 2004.

[17] R. Scha, "Taaltheorie en taaltechnologie; competence en performance," in *Computertoepassingen in de Neerlandistiek, LVVN-jaarboek*, 1990, pp. 7–22.

[18] D. Chiang, "Hierarchical phrase based translation," *Computational Linguistics, to appear*, 2007.

[19] A. Venugopal, A. Zollmann, and S. Vogel, "An efficient two-pass approach to synchronous CFG driven MT," in *Proc. of HLT/NAACL*, 2007.

[20] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of HLT/NAACL*, 2003.

[21] S. Vogel, H. Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation," in *Proc. of International Conference On Computational Linguistics*, 1996, pp. 836–841.

[22] N. A. Smith, D. A. Smith, and R. W. Tromble, "Context-based morphological disambiguation with random fields," in *HLT- EMNLP'05*, 2005, pp. 475–482.

[23] N. Habash and F. Sadat, "Morphological preprocessing scheme combination for statistical mt," in *Proceedings of COLING-ACL*, 2006.

[24] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.