

N-gram-based Statistical Machine Translation versus Syntax Augmented Machine Translation: comparison and system combination

Maxim Khalilov and José A.R. Fonollosa

Universitat Politècnica de Catalunya

Campus Nord UPC, 08034

Barcelona, Spain

{khalilov, adrian}@talp.upc.edu

Abstract

In this paper we compare and contrast two approaches to Machine Translation (MT): the CMU-UKA Syntax Augmented Machine Translation system (SAMT) and UPC-TALP N-gram-based Statistical Machine Translation (SMT). SAMT is a hierarchical syntax-driven translation system underlain by a phrase-based model and a target part parse tree. In N-gram-based SMT, the translation process is based on bilingual units related to word-to-word alignment and statistical modeling of the bilingual context following a maximum-entropy framework. We provide a step-by-step comparison of the systems and report results in terms of automatic evaluation metrics and required computational resources for a smaller Arabic-to-English translation task (1.5M tokens in the training corpus). Human error analysis clarifies advantages and disadvantages of the systems under consideration. Finally, we combine the output of both systems to yield significant improvements in translation quality.

1 Introduction

There is an ongoing controversy regarding whether or not information about the syntax of language can benefit MT or contribute to a hybrid system.

Classical IBM word-based models were recently augmented with a phrase translation capability, as shown in Koehn et al. (2003), or in more recent implementation, the MOSES MT system¹ (Koehn et al., 2007). In parallel to the phrase-based approach, the *N*-gram-based approach appeared (Mariño et al., 2006). It stems from

the Finite-State Transducers paradigm, and is extended to the log-linear modeling framework, as shown in (Mariño et al., 2006). A system following this approach deals with bilingual units, called tuples, which are composed of one or more words from the source language and zero or more words from the target one. The *N*-gram-based systems allow for linguistically motivated word reordering by implementing word order monotonicity.

Prior to the SMT revolution, a major part of MT systems was developed using rule-based algorithms; however, starting from the 1990's, syntax-driven systems based on phrase hierarchy have gained popularity. A representative sample of modern syntax-based systems includes models based on bilingual synchronous grammar (Melamed, 2004), parse tree-to-string translation models (Yamada and Knight, 2001) and non-isomorphic tree-to-tree mappings (Eisner, 2003).

The orthodox phrase-based model was enhanced in Chiang (2005), where a hierarchical phrase model allowing for multiple generalizations within each phrase was introduced. The open-source toolkit SAMT² (Zollmann and Venugopal, 2006) is a further evolution of this approach, in which syntactic categories extracted from the target side parse tree are directly assigned to the hierarchically structured phrases.

Several publications discovering similarities and differences between distinct translation models have been written over the last few years. In Crego et al. (2005b), the *N*-gram-based system is contrasted with a state-of-the-art phrase-based framework, while in DeNeeffe et al. (2007), the authors seek to estimate the advantages, weakest points and possible overlap between syntax-based MT and phrase-based SMT. In Zollmann et al. (2008) the comparison of phrase-based, "Chiang's style" hierarchical system and SAMT is pro-

¹www.statmt.org/moses/

²www.cs.cmu.edu/~zollmann/samt

vided.

In this study, we intend to compare the differences and similarities of the statistical N -gram-based SMT approach and the SAMT system. The comparison is performed on a small Arabic-to-English translation task from the news domain.

2 SAMT system

A criticism of phrase-based models is data sparseness. This problem is even more serious when the source, the target, or both languages are inflectional and rich in morphology. Moreover, phrase-based models are unable to cope with global reordering because the distortion model is based on movement distance, which may face computational resource limitations (Och and Ney, 2004).

This problem was successfully addressed when the MT system based on generalized hierarchically structured phrases was introduced and discussed in Chiang (2005). It operates with only two markers (a substantial phrase category and "a glue marker"). Moreover, a recent work (Zollmann and Venugopal, 2006) reports significant improvement in terms of translation quality if complete or partial syntactic categories (derived from the target side parse tree) are assigned to the phrases.

2.1 Modeling

A formalism for Syntax Augmented Translation is probabilistic synchronous context-free grammar (PSynCFG), which is defined in terms of source and target terminal sets and a set of non-terminals:

$$X \longrightarrow \langle \gamma, \alpha, \sim, \omega \rangle$$

where X is a non-terminal, γ is a sequence of source-side terminals and non-terminals, α is a sequence of target-side terminals and non-terminals, \sim is a one-to-one mapping from non-terminal tokens space in γ to non-terminal space in α , and ω is a non-negative weight assigned to the rule.

The non-terminal set is generated from the syntactic categories corresponding to the target-side Penn Treebank set, a set of glue rules and a special marker representing the "Chiang-style" rules, which do not span the parse tree. Consequently, all lexical mapping rules are covered by the phrases mapping table.

2.2 Rules annotation, generalization and pruning

The SAMT system is based on a purely lexical phrase table, which is identified as shown in

Koehn et al. (2003), and word alignment, which is generated by the *grow-diag-final-and* method (expanding the alignment by adding directly neighboring alignment points and alignment points in the diagonal neighborhood) (Och and Ney, 2003).

Meanwhile, the target of the training corpus is parsed with Charniak's parser (Charniak, 2000), and each phrase is annotated with the constituent that spans the target side of the rules. The set of non-terminals is extended by means of conditional and additive categories according to Combinatory Categorical Grammar (CCG) (Steedman, 1999). Under this approach, new rules can be formed. For example, $RB+VB$, can represent an additive constituent consisting of two synthetically generated adjacent categories³, i.e., an adverb and a verb. Furthermore, $DT \setminus NP$ can indicate an incomplete noun phrase with a missing determiner to the left.

The rule recursive generalization procedure coincides with the one proposed in Chiang (2005), but violates the restrictions introduced for single-category grammar; for example, rules that contain adjacent generalized elements are not discarded.

Thus, each rule

$$N \longrightarrow f_1 \dots f_m / e_1 \dots e_n$$

can be extended by another existing rule

$$M \longrightarrow f_i \dots f_u / e_j \dots e_v$$

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

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

where k is an index for the non-terminal M that indicates a one-to-one correspondence between the new M tokens on the two sides.

Figure 1 shows an example of initial rules extraction, which can be further extended using the hierarchical model, as shown in Figure 2 (consequently involving more general elements in rule description).

Rules pruning is necessary because the set of generalized rules can be huge. Pruning is performed according to the relative frequency and the nature of the rules: non-lexical rules that have been seen only once are discarded; source-conditioned rules with a relative frequency of appearance below a threshold are also eliminated.

³Adjacent generalized elements are not allowed in Chiang's work because of over-generation. However, over-generation is not an issue within the SAMT framework due to restrictions introduced by target-side syntax

Rules that do not contain non-terminals are not pruned.

2.3 Decoding and feature functions

The decoding process is accomplished using a top-down log-linear model. The source sentence is decoded and enriched with the PSynCFG in such a way that translation quality is represented by a set of feature functions for each rule, i.e.:

- rule *conditional probabilities*, given a source, a target or a left-hand-side category;
- *lexical weights features*, as described in Koehn et al. (2003);
- *counters* of target words and rule applications;
- *binary features* reflecting *rule context* (purely lexical and purely abstract, among others);
- rule *rareness* and *unbalancedness* penalties.

The decoding process can be represented as a search through the space of neg log probability of the target language terminals. The set of feature functions is combined with a finite-state target-side n-gram language model (LM), which is used to derive the target language sequence during a parsing decoding. The feature weights are optimized according to the highest BLEU score. For more details refer to Zollmann and Venugopal (2006).

3 UPC n-gram SMT system

A description of the UPC-TALP *N*-gram translation system can be found in Mariño et al. (2006).

SMT is based on the principle of translating a source sentence (*f*) into a sentence in the target language (*e*). The problem is formulated in terms of source and target languages; it is defined according to equation (1) and can be reformulated as selecting a translation with the highest probability from a set of target sentences (2):

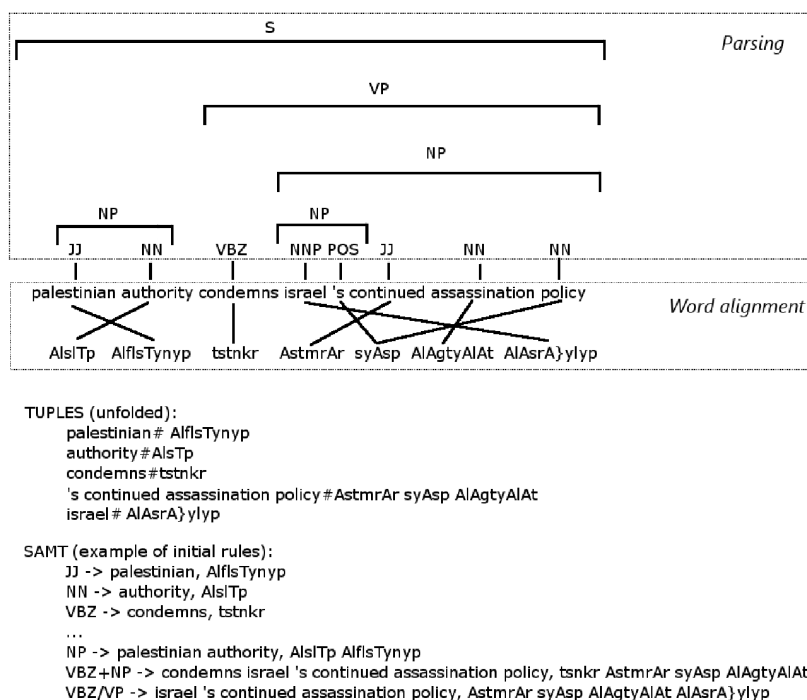


Figure 1: Example of SAMT and N-gram elements extraction.

SAMT generalized rules (example)

VBZ+NP -> VBZ israel 's continued assassina-tion policy, VBZ AstmrAr syAsp AlAgtyAlAt AlAsrA}ylyp
VBZ+NP -> VBZ NNP 's continued assassina-tion policy, VBZ AstmrAr syAsp AlAgtyAlAt NNP

Figure 2: Example of SAMT generalized rules.

$$\hat{e}_1^I = \arg \max_{e_1^I} \left\{ p(e_1^I | f_1^J) \right\} = \quad (1)$$

$$= \arg \max_{e_1^I} \left\{ p(f_1^J | e_1^I) \cdot p(e_1^I) \right\} \quad (2)$$

where I and J represent the number of words in the target and source languages, respectively.

Modern state-of-the-art SMT systems operate with the bilingual units extracted from the parallel corpus based on word-to-word alignment. They are enhanced by the *maximum entropy approach* and the posterior probability is calculated as a *log-linear combination* of a set of feature functions (Och and Ney, 2002). Using this technique, the additional models are combined to determine the translation hypothesis, as shown in (3):

$$\hat{e}_1^I = \arg \max_{e_1^I} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (3)$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

3.1 N-gram-based translation system

The N -gram approach to SMT is considered to be an alternative to the *phrase-based* translation, where a given source word sequence is decomposed into monolingual phrases that are then translated one by one (Marcu and Wong, 2002).

The N -gram-based approach regards translation as a stochastic process that maximizes the joint probability $p(f, e)$, leading to a decomposition based on bilingual n -grams. The core part of the system constructed in this way is a translation model (TM), which is based on bilingual units, called tuples, that are extracted from a word alignment (performed with GIZA++ tool⁴) according to certain constraints. A bilingual TM actually constitutes an n -gram LM of tuples, which approximates the joint probability between the languages under consideration and can be seen here as a LM, where the language is composed of tuples.

3.2 Additional features

The N -gram translation system implements a log-linear combination of five additional models:

- *an n -gram target LM;*

⁴<http://code.google.com/p/giza-pp/>

- *a target LM of Part-of-Speech tags;*
- *a word penalty model* that is used to compensate for the system's preference for short output sentences;
- *source-to-target and target-to-source lexicon models* as shown in Och and Ney (2004).

3.3 Extended word reordering

An extended monotone distortion model based on the automatically learned reordering rules was implemented as described in Crego and Mariño (2006). Based on the word-to-word alignment, tuples were extracted by an *unfolding* technique. As a result, the tuples were broken into smaller tuples, and these were sequenced in the order of the target words. An example of unfolding tuple extraction, contrasted with the SAMT chunk-based rules construction, is presented in Figure 1.

The reordering strategy is additionally supported by a 4-gram LM of reordered source POS tags. In training, POS tags are reordered according to the extracted reordering patterns and word-to-word links. The resulting sequence of source POS tags is used to train the n -gram LM.

3.4 Decoding and optimization

The open-source MARIE⁵ decoder was used as a search engine for the translation system. Details can be found in Crego et al. (2005a). The decoder implements a beam-search algorithm with pruning capabilities. All the additional feature models were taken into account during the decoding process. Given the development set and references, the log-linear combination of weights was adjusted using a *simplex* optimization method and an n -best re-ranking as described in <http://www.statmt.org/jhuws/>.

4 Experiments

4.1 Evaluation framework

As training corpus, we used the 50K first-lines extraction from the Arabic-English corpus that was provided to the NIST'08⁶ evaluation campaign and belongs to the news domain. The corpus statistics can be found in Table 1. The development and test sets were provided with 4 reference translations, belong to the same domain and contain 663 and 500 sentences, respectively.

⁵<http://gps-tsc.upc.es/veu/soft/soft/marie/>

⁶www.nist.gov/speech/tests/mt/2008/

	Arabic	English
Sentences	50 K	50 K
Words	1.41 M	1.57 K
Average sentence length	28.15	31.22
Vocabulary	51.10 K	31.51 K

Table 1: Basic statistics of the training corpus.

Evaluation conditions were case-insensitive and sensitive to tokenization. The word alignment is automatically computed by using GIZA++ (Och and Ney, 2004) in both directions, which are made symmetric by using the *grow-diag-final-and* operation.

The experiments were done on a dual-processor Pentium IV Intel Xeon Quad Core X5355 2.66 GHz machine with 24 G of RAM. All computational times and memory size results are approximated.

4.2 Arabic data preprocessing

Arabic is a VSO (SVO in some cases) pro-drop language with rich templatic morphology, where words are made up of roots and affixes and clitics agglutinate to words. For preprocessing, a similar approach to that shown in Habash and Sadat (2006) was employed, and the MADA+TOKAN system for disambiguation and tokenization was used. For disambiguation, only diacritic unigram statistics were employed. For tokenization, the D3 scheme with -TAGBIES option was used. The scheme splits the following set of clitics: w+, f+, b+, k+, l+, Al+ and pronominal clitics. The -TAGBIES option produces Bies POS tags on all taggable tokens.

4.3 SAMT experiments

The SAMT guideline was used to perform the experiments and is available on-line: <http://www.cs.cmu.edu/~zollmann/samt/>.

Moses MT script was used to create the *grow - diag - final* word alignment and extract purely lexical phrases, which are then used to induce the SAMT grammar. The target side (English) of the training corpus was parsed with the Charniak’s parser (Charniak, 2000).

Rule extraction and filtering procedures were restricted to the concatenation of the development and test sets, allowing for rules with a maximal length of 12 elements in the source side and with a

zero minimum occurrence criterion for both non-lexical and purely lexical rules.

Moses-style phrases extracted with a phrase-based system were 4.8M, while a number of generalized rules representing the hierarchical model grew dramatically to 22.9M. 10.8M of them were pruned out on the filtering step.

The vocabulary of the English Penn Treebank elementary non-terminals is 72, while a number of generalized elements, including additive and truncated categories, is 35.7K.

The *FastTranslateChart* beam-search decoder was used as an engine of MER training aiming to tune the feature weight coefficients and produce final n-best and 1-best translations by combining the intensive search with a standard 4-gram LM as shown in Venugopal et al. (2007). The iteration limit was set to 10 with 1000-best list and the highest BLEU score as optimization criteria. We did not use completely abstract rules (without any source-side lexical utterance), since these rules significantly slow down the decoding process (*noAllowAbstractRules* option).

Table 2 shows a summary of computational time and RAM needed at each step of the translation.

Step	Time	Memory
Parsing	1.5h	80Mb
Rules extraction	10h	3.5Gb
Filtering&merging	3h	4.0Gb
Weights tuning	40h	3Gb
Testing	2h	3Gb

Table 2: SAMT: Computational resources.

Evaluation scores including results of system combination (see subsection 4.6) are reported in Table 3.

4.4 N-gram system experiments

The core model of the *N*-gram-based system is a 4-gram LM of bilingual units containing: 184.345 1-grams⁷, 552.838 2-grams, 179.466 3-grams and 176.221 4-grams.

Along with this model, an *N*-gram SMT system implements a log-linear combination of a 5-gram target LM estimated on the English portion of the parallel corpus, as well as supporting 4-gram source and target models of POS tags. *Bies*

⁷This number also corresponds to the bilingual model vocabulary.

	BLEU	NIST	mPER	mWER	METEOR
SAMT	43.20	9.26	36.89	49.45	58.50
N-gram-based SMT	46.39	10.06	32.98	48.47	62.36
System combination	48.00	10.15	33.20	47.54	62.27
MOSES Factored System	44.73	9.62	33.92	47.23	59.84
Oracle	61.90	11.41	28.84	41.52	66.19

Table 3: Test set evaluation results

POS tags were used for the Arabic portion, as shown in subsection 4.2; a *TnT* tool was used for English POS tagging (Brants, 2000).

The number of non-unique initially extracted tuples is $1.1M$, which were pruned according to the maximum number of translation options per tuple on the source side (30). Tuples with a NULL on the source side were attached to either the previous or the next unit (Mariño et al., 2006). The feature models weights were optimized according to the same optimization criteria as in the SAMT experiments (the highest BLEU score).

Stage-by-stage RAM and time requirements are presented in Table 4, while translation quality evaluation results can be found in Table 3.

Step	Time	Memory
Models estimation	0.2h	1.9Gb
Reordering	1h	—
Weights tuning	15h	120Mb
Testing	2h	120Mb

Table 4: Tuple-based SMT: Computational resources.

4.5 Statistical significance

A statistical significance test based on a bootstrap resampling method, as shown in Koehn (2004), was performed. For the 98% confidence interval and 1000 set resamples, translations generated by SAMT and *N*-gram system are significantly different according to BLEU (43.20 ± 1.69 for SAMT vs. 46.42 ± 1.61 for tuple-based system).

4.6 System combination

Many MT systems generate very different translations of similar quality, even if the models involved into translation process are analogous. Thus, the outputs of syntax-driven and purely statistical MT systems were combined at the sentence level using 1000-best lists of the most probable

translations produced by the both systems.

For system combination, we followed a Minimum Bayes-risk algorithm, as introduced in Kumar and Byrne (2004). Table 3 shows the results of the system combination experiments on the test set, which are contrasted with the *oracle* translation results, performed as a selection of the translations with the highest BLEU score from the union of two 1000-best lists generated by SAMT and *N*-gram SMT.

We also analyzed the percentage contribution of each system to the system combination: 55-60% of best translations come from the *tuples*-based system 1000-best list, both for system combination and oracle experiments on the test set.

4.7 Phrase-based reference system

In order to understand the obtained results compared to the state-of-the-art SMT, a reference phrase-based factored SMT system was trained and tested on the same data using the MOSES toolkit. *Surface* forms of words (factor “0”), *POS* (factor “1”) and canonical forms of the words (*lemmata*) (factor “2”) were used as English factors, and *surface* forms and *POS* were the Arabic factors.

Word alignment was performed according to the *grow-diag-final* algorithm with the GIZA++ tool, a *msd-bidirectional-fe* conditional reordering model was trained; the system had access to the target-side 4-gram LMs of words and POS. The $0-0, 1+0-1, 2+0-1$ scheme was used on the translation step and $1, 2-0, 1+1-0, 1$ to create generation tables. A detailed description of the model training can be found on the MOSES tutorial web-page⁸. The results may be seen in Table 3.

5 Error analysis

To understand the strong and weak points of both systems under consideration, a human analysis of

⁸<http://www.statmt.org/moses/>

the typical translation errors generated by each system was performed following the framework proposed in Vilar et al. (2006) and contrasting the systems output with four reference translations. Human evaluation of translation output is a time-consuming process, thus a set of 100 randomly chosen sentences was picked out from the corresponding system output and was considered as a representative sample of the automatically generated translation of the test corpus. According to the proposed error topology, some classes of errors can overlap (for example, an unknown word can lead to a reordering problem), but it allows finding the most prominent source of errors in a reliable way (Vilar et al., 2006; Povovic et al., 2006). Table 5 presents the comparative statistics of errors generated by the SAMT and the N -gram-based SMT systems. The average length of the generated translations is 32.09 words for the SAMT translation and 35.30 for the N -gram-based system.

Apart from unknown words, the most important sources of errors of the SAMT system are missing content words and extra words generated by the translation system, causing 17.22 % and 10.60 % of errors, respectively. A high number of missing content words is a serious problem affecting the translation accuracy. In some cases, the system is able to construct a grammatically correct

translation, but omitting an important content word leads to a significant reduction in translation accuracy:

SAMT translation: *the ministers of arab environment for the closure of the Israeli dymwnp reactor .*

Ref 1: *arab environment ministers demand the closure of the Israeli daemona nuclear reactor .*

Ref 2: *arab environment ministers demand the closure of Israeli dimona reactor .*

Ref 3: *arab environment ministers call for Israeli nuclear reactor at dimona to be shut down .*

Ref 4: *arab environmental ministers call for the shutdown of the Israeli dimona reactor .*

Extra words embedded into the correctly translated phrases are a well-known problem of MT systems based on hierarchical models operating on the small corpora. For example, in many cases the Arabic expression AlbHr Almyt is translated into English as dead sea side and not as dead sea, since the bilingual instances contain only the whole English phrase, like following:

AlbHr Almyt#the dead sea side#@NP

The N -gram-based system handles missing words more correctly – only 9.40 % of the errors come from the missing content

Type	Sub-type	SAMT	N-gram
Missing words		152 (25.17 %)	92 (15.44 %)
	Content words	104 (17.22 %)	56 (9.40 %)
	Filler words	48 (7.95 %)	36 (6.04 %)
Word order		96 (15.89 %)	140 (23.49 %)
	Local word order	20 (3.31 %)	68 (11.41 %)
	Local phrase order	20 (3.31 %)	20 (3.36 %)
	Long range word order	32 (5.30 %)	48 (8.05 %)
	Long range phrase order	24 (3.97 %)	4 (0.67 %)
Incorrect words		164 (27.15 %)	204 (34.23 %)
	Sense: wrong lexical choice	24 (3.97 %)	60 (10.07 %)
	Sense: incorrect disambiguation	16 (2.65 %)	8 (1.34 %)
	Incorrect form	24 (3.97 %)	56 (9.40 %)
	Extra words	64 (10.60 %)	56 (9.40 %)
	Style	28 (4.64 %)	20 (3.36 %)
	Idioms	4 (0.07 %)	4 (0.67 %)
Unknown words		132 (21.85 %)	104 (17.45 %)
Punctuation		60 (9.93 %)	56 (9.40 %)
Total		604	596

Table 5: Human made error statistics for a representative test set.

words; however, it does not handle local and long-term reordering, thus the main problem is phrase reordering (11.41 % and 8.05 % of errors). In the example below, the underlined block (Circumstantial Complement: from local officials in the tourism sector) is embedded between the verb and the direct object, while in correct translation it must be placed in the end of the sentence.

N-gram translation: *the winner received from local officials in the tourism sector three gold medals .*

Ref 1: *the winner received three gold medals from local officials from the tourism sector .*

Ref 2: *the winner received three gold medals from the local tourism officials .*

Ref 3: *the winner received his prize of 3 gold medals from local officials in the tourist industry .*

Ref 4: *the winner received three gold medals from local officials in the tourist sector .*

Along with inserting extra words and wrong lexical choice, another prominent source of incorrect translation, generated by the N -gram system, is an erroneous grammatical form selection, i.e., a situation when the system is able to find the correct translation but cannot choose the correct form. For example, arab environment minister call for closing dymwnp Israeli reactor, where the verb-preposition combination call for was correctly translated on the stem level, but the system was not able to generate a third person conjugation calls for. In spite of the fact that English is a language with nearly no inflection, 9.40 % of errors stem from poor word form modeling. This is an example of the weakest point of the SMT systems having access to a small training material; the decoder does not use syntactic information about the subject of the sentence (singular) and makes a choice only concerning the tuple probability.

The difference in total number of errors is negligible, however a subjective evaluation of the systems output shows that the translation generated by the N -gram system is more understandable than the SAMT one, since more content words are translated correctly and the meaning of the sentence is still preserved.

6 Discussion and conclusions

In this study two systems are compared: the UPC-TALP N -gram-based and the CMU-UKA SAMT systems, originating from the ideas of Finite-State Transducers and hierarchical phrase translation, respectively. The comparison was created to be as fair as possible, using the same training material and the same tools on the preprocessing, word-to-word alignment and language modeling steps. The obtained results were also contrasted with the state-of-the-art phrase-based SMT.

Analyzing the automatic evaluation scores, the N -gram-based approach shows good performance for the small Arabic-to-English task and significantly outperforms the SAMT system. The results shown by the modern phrase-based SMT (factored MOSES) lie between the two systems under consideration. Considering memory size and computational time, the tuple-based system has obtained significantly better results than SAMT, primarily because of its smaller search space.

Interesting results were obtained for the PER and WER metrics: according to the PER, the UPC-TALP system outperforms the SAMT by 10%, while the WER improvement hardly achieves a 2% difference. The N -gram-based SMT can translate the context better, but produces more reordering errors than SAMT. This may be explained by the fact that Arabic and English are languages with high disparity in word order, and the N -gram system deals worse with long-distance reordering because it attempts to use shorter units. However, by means of introducing the word context into the TM, short-distance bilingual dependencies can be captured effectively.

The main conclusion that can be made from the human evaluation analysis is that the systems commit a comparable number of errors, but they are distributed dissimilarly. In case of the SAMT system, the frequent errors are caused by missing or incorrectly inserted extra words, while the N -gram-based system suffers from reordering problems and wrong words/word form choice

Significant improvement in translation quality was achieved by combining the outputs of the two systems based on different translating principles.

7 Acknowledgments

This work has been funded by the Spanish Government under grant TEC2006-13964-C03 (AVI-VAVOZ project).

References

- T. Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000)*.
- E. Charniak. 2000. A maximum entropy-inspired parser. In *Proceedings of NAACL 2000*, pages 132–139.
- D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.
- J. M. Crego and J. B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.
- J. M. Crego, J. Mariño, and A. de Gispert. 2005a. An Ngram-based Statistical Machine Translation Decoder. In *Proceedings of INTERSPEECH05*, pages 3185–3188.
- J.M. Crego, M.R. Costa-jussà, J.B. Mariño, and J.A.R. Fonollosa. 2005b. Ngram-based versus phrase-based statistical machine translation. In *Proc. of the IWSLT 2005*, pages 177–184.
- S. DeNeefe, K. Knight, W. Wang, and D. Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proceedings of EMNLP-CoNLL 2007*, pages 755–763.
- J. Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of ACL 2003 (companion volume)*, pages 205–208.
- N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of HLT/NAACL 2006*, pages 49–52.
- Ph. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based machine translation. In *Proceedings of HLT-NAACL 2003*, pages 48–54.
- Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open-source toolkit for statistical machine translation. In *Proceedings of ACL 2007*, pages 177–180.
- P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395.
- S. Kumar and W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT/NAACL 2004*.
- D. Marcu and W. Wong. 2002. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of EMNLP02*, pages 133–139.
- J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.
- I.D. Melamed. 2004. Statistical machine translation by parsing. In *Proceedings of ACL 2004*, pages 111–114.
- F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of ACL 2002*, pages 295–302.
- F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.
- M. Povovic, A. de Gispert, D. Gupta, P. Lambert, J.B. Mariño, M. Federico, H. Ney, and R. Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *In Proceeding of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 1–6.
- M. Steedman. 1999. Alternative quantifier scope in ccg. In *Proceedings of ACL 1999*, pages 301–308.
- A. Venugopal, A. Zollmann, and S. Vogel. 2007. An Efficient Two-Pass Approach to Synchronous-CFG Driven Statistical MT. In *Proceedings of HLT/NAACL 2007*, pages 500–507.
- D. Vilar, J. Xu, L. F. D’Haro, and H. Ney. 2006. Error Analysis of Machine Translation Output. In *Proceedings of LREC’06*, pages 697–702.
- K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL 2001*, pages 523–530.
- A. Zollmann and A. Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of NAACL 2006*.
- A. Zollmann, A. Venugopal, F. Och, and J. Ponte. 2008. Systematic comparison of Phrase-based, Hierarchical and Syntax-Augmented Statistical mt. In *Proceedings of Coling 2008*, pages 1145–1152.