Expanding the Language model in a low-resource hybrid MT system

George Tambouratzis
ILSP, Athena R.C
giorg_t@ilsp.gr

Sokratis Sofianopoulos
ILSP, Athena R.C
s_sofian@ilsp.gr

Marina Vassiliou
ILSP, Athena R.C
mvas@ilsp.gr

Abstract

The present article investigates the fusion of different language models to improve translation accuracy. A hybrid MT system, recently-developed in the European Commission-funded PRESEMT project that combines example-based MT and Statistical MT principles is used as a starting point. In this article, the syntactically-defined phrasal language models (NPs, VPs etc.) used by this MT system are supplemented by n-gram language models to improve translation accuracy. For specific structural patterns, n-gram statistics are consulted to determine whether the pattern instantiations are corroborated. Experiments indicate improvements in translation accuracy.

1 Introduction

Currently a major part of cutting-edge research in MT revolves around the statistical machine translation (SMT) paradigm. SMT has been inspired by the use of statistical methods to create language models for a number of applications including speech recognition. A number of different translation models of increasing complexity and translation accuracy have been developed (Brown et al., 1993). Today, several packages for developing statistical language models are available for free use, including SRI (Stolke et al., 2011), thus supporting research into statistical methods. A main reason for the widespread adoption of SMT is that it is directly amenable to new language pairs using the same algorithms. An integrated framework (MOSES) has been developed for the creation of SMT systems (Koehn et al., 2007). The more recent developments of SMT are summarised by Koehn (2010). One particular advance in SMT has been the integration of syntactically motivated phrases in order to establish correspondences between source language (SL) and target language (TL) (Koehn et al., 2003). Recently SMT has been enhanced by using different levels of abstraction e.g. word, lemma or part-of-speech (PoS), in factored SMT models so as to improve SMT performance (Koehn & Hoang, 2007).

The drawback of SMT is that SL-to-TL parallel corpora of the order of millions of tokens are required to extract meaningful models for translation. Such corpora are hard to obtain, particularly for less resourced languages. For this reason, SMT researchers are increasingly investigating the extraction of information from monolingual corpora, including lexica (Koehn & Knight, 2002 & Klementiev et al., 2012), restructuring (Nuhn et al., 2012) and topic-specific information (Su et al., 2011).

As an alternative to pure SMT, the use of less specialised but more readily available resources has been proposed. Even if such approaches do not provide a translation quality as high as SMT, their ability to develop MT systems with very limited resources confers to them an important advantage. Carbonell et al. (2006) have proposed an MT method that requires no parallel text, but relies on a full-form bilingual dictionary and a decoder using long-range context. Other systems using low-cost resources include METIS (Dologlou et al., 2003) and METIS-II (Markantonatou et al., 2009), which are based only on large monolingual corpora to translate SL texts.

Another recent trend in MT has been towards hybrid MT systems, which combine characteristics from multiple MT paradigms. The idea is that by fusing characteristics from different paradigms, a better translation performance can be attained (Wu et al., 2005). In the present article, the PRESEMT hybrid MT method using predominantly monolingual corpora (Sofianopoulos et al., 2012 & Tambouratzis et al., 2013) is extended by integrating n-gram information to improve the translation accuracy. The focus of the article is on how to extract, as comprehensively as possible, information from monolingual corpora by combining multiple models, to allow a higher quality translation.

A review of the base MT system is performed in section 2. The TL language model is then detailed, allowing new work to be presented in section 3. More specifically, via an error analysis, n-gram based extensions are proposed to augment

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 57–66,
October 25, 2014, Doha, Qatar.

the language model. Experiments are presented in section 4 and discussed in section 5.

2 The hybrid MT methodology in brief

The PRESEMT methodology can be broken down into the pre-processing stage, the post-processing stage and two translation steps each of which addresses different aspects of the translation process. The first translation step establishes the structure of the translation by performing a structural transformation of the source side phrases based on a small bilingual corpus, to capture long range reordering. The second step makes lexical choices and performs local word reordering within each phrase. By dividing the translation process in these two steps the challenging task of both local and long distance reordering is addressed.

Phrase-based SMT systems give accurate translations for language pairs that only require a limited number of short-range reorderings. On the contrary, when translating between languages with free word order, these models prove inefficient. Instead, reordering models need to be built, which require large parallel training data, as various reordering challenges must be tackled.

2.1 Pre-processing

This involves PoS tagging, lemmatising and shallow syntactic parsing (chunking) of the source text. In terms of resources, the methodology utilises a bilingual lemma dictionary, an extensive TL monolingual corpus, annotated with PoS tags, lemmas and syntactic phrases (chunks), and a very small parallel corpus of 200 sentences, with tagged and lemmatised source side and tagged, lemmatised and chunked target side. The bilingual corpus provides samples of the structural transformation from SL to TL. During this phase, the translation methodology ports the chunking from the TL- to the SL-side, alleviating the need for an additional parser in SL. An example of the pre-processing stage is shown in Figure 1, for a sentence translated from Greek to English. For this sentence, the chunk structure is shown at the bottom part of Figure 1.

2.2 Structure Selection

Structure selection transforms the input text using the limited bilingual corpus as a structural knowledge base, closely resembling the "translation by analogy" aspect of EBMT systems (Hutchins, 2005). Using available structural information, namely the order of syntactic phrases, the PoS tag of the head token of each phrase and the case of the head token (if available), we retrieve the most similar source side sentence from the parallel corpus. Based on the alignment information from the bilingual corpus between SL and TL, the input sentence structure is transformed to the structure of the target side translation.

For the retrieval of the most similar source side sentence, an algorithm from the dynamic programming paradigm is adopted (Sofianopoulos et al., 2012), treating the structure selection process as a sequence alignment, aligning the input sentence to an SL side sentence from the aligned parallel corpus and assigning a similarity score. The implementation is based on the Smith-Waterman algorithm (Smith and Waterman, 1981), initially proposed for similarity detection between protein sequences. The algorithm finds the optimal local alignment between the two input sequences at clause level.

The similarity of two clauses is calculated by taking into account the edit operations (replacement, insertion or removal) that must be applied to the input sentence in order to transform it to a source side sentence from the corpus. Each of these operations has an associated cost, considered as a system parameter. The parallel corpus sentence that achieves the highest similarity score is the most similar one to the input source sentence. For the example of Figure 1, the comparison of the SL sentence structure to the parallel corpus is schematically depicted in Figure 2. The resulting TL sentence structure is shown in Figure 3 in terms of phrase types and heads.

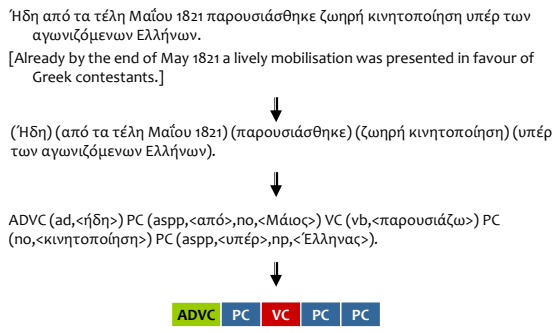

Figure 1. Pre-processing of sentence (its gloss in square brackets) into a chunk sequence.

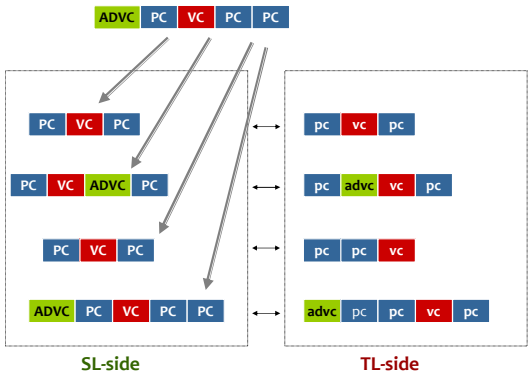

Figure 2. Comparing sentence structure to parallel corpus templates, to determine the best-matching SL structure (here, the 4th entry).

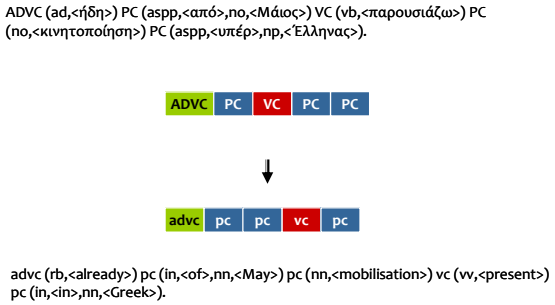

Figure 3. SL-to-TL Structure transformation based on the chosen parallel corpus template.

2.3 Translation equivalent selection

This second translation step performs word translation disambiguation, local word reordering within each syntactic phrase as well as addition and/or deletion of auxiliary verbs, articles and prepositions. All of the above are performed by using a syntactic phrase model extracted from a purely monolingual TL corpus. The final translation is produced by the token generation component, since all processing during the translation process is lemma-based.

Each sentence contained within the text to be translated is processed separately, so there is no exploitation of inter-sentential information. The first task is to select the correct TL translation of each word. The second task involves establishing the correct word order within each phrase. For each phrase of the sentence being translated, the algorithm searches the TL phrase model for similar phrases. All retrieved TL phrases are compared to the phrase to be translated. The comparison is based on the words included, their tags and lemmas and any other morphological features (case, number etc.). The stable-marriage algorithm (Gale & Shapley, 1962) is applied for calculating the similarity and aligning the words of a phrase pair.

This word reordering process is performed simultaneously with the translation disambiguation, using the same TL phrase model. During word reordering the algorithm also resolves issues regarding the insertion or deletion of articles and other auxiliary tokens. Though translation equivalent selection implements several tasks simultaneously, it produces encouraging results when translating from Greek (a free-word order language) to English (an SVO language).

2.4 Post-processing

In this stage, a token generator is applied to the lemmas of the translated sentences together with the morphological features of their equivalent source words, to produce the final word forms.

2.5 Comparison of the method to SMT

In the proposed methodology, the structure selection step performs long distance reordering without resorting to syntactic parsers and without employing any rules. In phrase-based SMT, long distance reordering is performed by either using SL syntax, with the use of complex reordering rules, or by using syntactic trees.

The similarity calculation algorithms used in the two translation steps of the proposed method are of a similar nature to the extraction of translation models in factored-based SMT. In SMT, different matrices are created for each model (i.e. one for lemmas and another one for PoS tags), while in the methodology studied here lemmas and tags are handled at the same time.

The main advantage of the method studied here is its ability to create a functioning MT system with a parallel corpus of only a few sentences (200 sentences in the present experiments). On the contrary, it would not be possible to create a working SMT with such a corpus.

3 Information extraction from the monolingual corpus

3.1 Standard indexed phrase model

The TL monolingual corpus is processed to extract two complementary types of information, both employed at the second phase of the translation process (cf. sub-section 2.3). The first implements a disambiguation between multiple possible translations, while the second provides the micro-structural information to establish token order in the final translation.

Both these types of information are extracted from one model. More specifically, during preprocessing of the corpus, a phrase model is established that provides the micro-structural information on the translation output, to determine intra-phrasal word order. The model is stored in a file structure, where a separate file is created for phrases according to their (i) type, (ii) head and (iii) head PoS tag.

The TL phrases are then organised in a hash map that allows the storage of multiple values for each key, using as a key the three aforementioned criteria. For each phrase the number of occurrences within the corpus is also retained. Each hash map is stored independently in a file for very fast access by the search algorithm. As a result of this process hundreds of thousands of files are generated, one for each combination of the three aforementioned criteria. Each file is of a small size and thus can be retrieved quickly.

For creating the model used here, a corpus of 30,000 documents has been processed for the TL, where each document contains a concatenation of independent texts of approximately 1MByte in size. The resulting phrase model consists of 380,000 distinct files, apportioned into 12,000 files of adjectival chunks, 348,000 of noun chunks, 17,000 of verb chunks and 3,000 of adverbial chunks. A sample of the indexed file corresponding to verb phrases with head 'help' is shown in Figure 4.

	Occurrences	Phrase structure
1	41448	help (VV)
2	29575	to(TO) help(VV)
3	5896	will(MD) help(VV)
4	4795	can(MD) help(VV)
5	2632	have(VHD) help(VVN)

Figure 4. Example of indexed file for "help".

3.2 Error analysis on translation output

In Table 1, the translation accuracy attained by the proposed hybrid approach in comparison to established systems is displayed. The proposed method occupies the middle ground between the two higher performing SMT-based systems (Bing and Google) and the Systran and WorldLingo commercial systems.

Though the BLEU score of the proposed method is 0.17 BLEU points lower than the Google score, the proposed method achieves what is a respectable score with a parallel corpus of only 200 sentences. Though the exact resources for Google or Bing are not disclosed, it is widely agreed that they are at least 3 orders of magnitude larger (very likely even more) justifying the lower scores achieved by the proposed low-resource method.

Number of sentences	200	Resources	stand.
Reference translations	1	Language pair	EL–EN

MT config.	Metrics			
	BLEU	NIST	Meteor	TER
PRESEMT-baseline	0.3462	6.974	0.3947	51.05
Google	0.5259	8.538	0.4609	42.23
Bing	0.4974	8.279	0.4524	34.18
SYSTRAN	0.2930	6.466	0.3830	49.72
WorldLingo	0.2659	5.998	0.3666	50.63

Table 1. Values of performance metrics for dataset1, using the baseline version of the proposed method and other established systems.

The n-gram method proposed in this article for supplementary language modelling is intended to identify recurring errors in the output or to verify translation choices made by the indexed monolingual model. The errors mainly concern generation of tokens out of lemmata, positioning of tokens within phrases as well as disambiguation choices. An indicative list of errors encountered for Greek to English translation follows:

Article introduction & deletion: Given that there is no 1:1 mapping between Greek and English concerning the use of the definite article, it is essential to check whether it is correctly introduced in specific cases (e.g. before proper names).

Generation of verb forms: Specific errors of the MT system involve cases of active/passive voice mismatches between SL and TL and deponent verbs, i.e. active verbs with mediopassive morphology. For example, the Greek deponent verb "έρχομαι" (come) is translated to "be come" by the system token generation component that takes into account the verb's passive morphology in SL. This erroneous translation should be corrected to "come", i.e. the auxiliary verb "be" must be deleted.

In-phrase token order: The correct ordering of tokens within a given phrase (which occasionally fails to be established by the proposed system) can be verified via the n-gram model.

Prepositional complements: When translating the prepositional complement of a verb (cf. "depend + on"), it is often the case that the incorrect preposition is selected during disambiguation, given that no context information is avail-

able. The n-gram model may be accessed to identify the appropriate preposition.

Double preposition: Prepositions appearing in succession within a sentence need to be reduced to one. For instance, the translation of the NP "*κατά τη διάρκεια της πολιορκίας*" (= during the siege) results in a prepositional sequence ("*during of*") due to the translation of the individual parts as follows:

κατά τη διάρκεια = **during**
της = **of** the
πολιορκίας = siege

In this example a single preposition is needed.

3.3 Introducing n-gram models

A new model based on n-gram appearances is intended to supplement phrase-based information already extracted from the monolingual corpus (cf. section 3.1). As the monolingual corpus is already lemmatised, both lemma and token-based n-grams are extracted. To simplify processing, no phrase-boundary information is retained in the n-gram models.

One issue is how the n-gram model will be combined with the indexed phrase model of the hybrid MT algorithm. The new n-gram model can be applied at the same stage of the translation process. Alternatively, n-grams can be applied after the indexed phrase model, for verification or revision of the translation produced by using the indexed corpus. Then, the indexed phrase model generates a first translation, which represents a hypothesis H_i, upon which a number of tests are performed. If the n-gram model corroborates this hypothesis, no modification is applied, whilst if the n-gram likelihood estimates lead to the rejection of the hypothesis, the translation is revised accordingly.

Having adopted this set-up, the main task is to specify the hypotheses to be tested. To that end, a data-driven approach based on the findings of the error analysis (cf. section 3.2) is used.

The creation of the TL n-gram model is straightforward and employs the publicly available SRILM tool (Stolke et al., 2011) to extract n-gram probabilities. Both 2-gram and 3-gram models have been extracted, creating both token-based and lemma-based models to support queries in factored representation levels. The n-gram models have used 20,000 documents in English, each document being an assimilation of web-posted texts with a cumulative size of 1 Mbyte (harvested without any restrictions in terms of domain). Following a pre-processing to remove words with non-English characters, the final corpus contains a total of 707.6 million tokens and forms part of the EnTenTen corpus1. When creating both 2-grams and 3-grams, Witten-Bell smoothing is used and all n-grams with less than 5 occurrences are filtered out to reduce the model size. Each n-gram model contains circa 25 million entries, which are the SRILM-derived logarithms of probabilities.

3.4 Establishing translation hypotheses

A set of hypotheses has been established based on the error analysis, to improve the translation quality. Each hypothesis is expressed by a mathematical formula which checks the likelihood of an n-gram, via either the lemma-based n-gram model (the relevant entry being denoted as *p_lem()*, i.e. the probability of the n-gram of lemmas) or the token-based model (the relevant entry being denoted as *p_tok*). The relevant 2-gram or 3-gram model is consulted depending on whether the number of arguments is 2 or 3.

Hypothesis H_1: This hypothesis checks for the existence of a deponent verb, i.e. verb which is in passive voice in SL but has an active voice translation. Instead of externally providing a list of deponent verbs in Greek, the n-gram model is used to determine translations for which the verb is always in active voice, by searching the frequency-of-occurrence in the TL corpus. As an example of a correct rejection of hypothesis H_1, consider the verb "κοιμάμαι" [to sleep] which is translated by the hybrid MT system into "be slept" as in SL this verb has a medio-passive morphology. As the pattern "be slept" is extremely infrequent in the monolingual corpus, hypothesis H_1 is rejected and lemma "be" is correctly deleted, to translate "κοιμάμαι" into "sleep". The corresponding hypothesis is:

$$H_1 : p\_lem(A,B) > thres\_h1,$$
$$\text{where } Lem(A) = \text{"be" and } PoS(B) = \text{"VVN"}$$

If the aforementioned hypothesis does not hold, (i.e. the probability of the 2-gram formed by the auxiliary verb with lemma B is very rare) then H_1 is rejected and the auxiliary verb is deleted, as expressed by the following formula:

$$If\ (H_1 == false)\ then\ \{A, B\} \rightarrow \{B\}$$

Hypothesis H_2: This hypothesis checks the inclusion of an article, within a trigram of word forms. If this hypothesis is rejected based on n-gram evidence, the article is deleted. Hypothesis

[1] http://www.sketchengine.co.uk/documentation/wiki/Corpora/enTenTen

H_2 is expressed as follows, where *thres_h2* is a minimum threshold margin:

$$H_2: min\{p\_lem(A,the),p\_lem(the,B)\} - p\_lem(A;B) < thres\_h_2$$

An example of correctly rejecting H_2 is for trigram {see, the, France}, which is revised to {see, France}.

$$If\ (H_2 == false)\ then\ \{A, the, B\} \rightarrow \{A, B\}$$

Hypothesis H_3: This hypothesis is used to handle cases where two consecutive prepositions exist (for prepositions the PoS tag is "IN"). In this case one of these prepositions must be deleted, based on the n-gram information. This process is expressed as follows:

$$H_3 : max((p\_lem(A;B),p\_lem(A,C)),\ where\ PoS(A)==\text{"IN"}\ \&\ PoS(B)==\text{"IN"}$$

$$If\ (H_3==TRUE)\ then\ \{A, B, C\} \rightarrow \{A, C\}\ or\ \{B, C\}$$

Hypothesis H_4: This hypothesis checks if there exists a more suitable preposition than the one currently selected for a given trigram {A, B, C}, where PoS(B) = "IN". H_4 is expressed as:

$$H_4: p\_lem(A,B,C)-max(p\_lem(A,D,C)>thres\_h4\ ,\ for\ all\ D\ where\ PoS\{D\}==\text{"IN"}.$$

If this hypothesis is rejected, B is replaced by D:

$$If\ (H_4==FALSE)\ then\ (\{A,B,C\} \rightarrow \{A,D,C\}$$

Hypothesis H_5: This hypothesis checks if for a bigram, the wordforms might be replaced by the corresponding lemmas, as the wordform-based pattern is too infrequent. This is formulated as:

$$H_5: p\_tok(A,B)- p\_tok(lem(A),lem(B)) > thres\_h5$$

An example application would involve processing bigram {can, is} and revising it into the correct {can, be} by rejecting H5:

$$If\ (H_5==FALSE)\ then\ \{A,B\ \} \rightarrow \{lem(A),lem(B)\}$$

Similarly, H5 can revise the plural form "informations" to the correct "information".

Hypothesis H_6: This hypothesis also handles article deletion, by studying however bigrams, rather than trigrams, (cf. H_1). This hypothesis is that the bigram frequency exceeds a given threshold value (*thres_6*).

$$H_6 :p\_lem(2\text{-}gram(A, B))>thres\_h6,\ where\ PoS(A)=\text{"DT"}$$

If H_6 is rejected, the corresponding article is deleted, as indicated by the following formula:

$$If\ (H_6==FALSE)\ then\ \{A,B\} \rightarrow \{B\}$$

4 Objective Evaluation Experiments

4.1 Experiment design

The experiments reported in the present article focus on the Greek – English language pair, the reason being that this is the language pair for which the most extensive experimentation has been reported for the PRESEMT system (Tambouratzis et al., 2013). Thus, improvements in the translation accuracy will be more difficult to attain. Two datasets are used to evaluate translation accuracy, a development set (dataset1) and a test set (dataset2), each containing 200 sentences of length ranging from 7 to 40 tokens. These sets of sentences are readily available for download over the project website[2]. Two versions of the bilingual lexicon have been used, a base version and an expanded one.

Both sets are manually translated by Greek native speakers and then cross-checked by English native speakers, with one reference translation per sentence. A range of evaluation metrics are employed, namely BLEU (Papineni et al., 2002), NIST (NIST 2002), Meteor (Denkowski and Lavie, 2011) and TER (Snover et al., 2006).

4.2 Experimental results

The exact sequence with which hypotheses are tested affects the results of the translation, since only one hypothesis is allowed to be applied to each sentence token at present. This simplifies the evaluation of the hypotheses' effectiveness. As a result, hypotheses are applied in strict order (i.e. first H_1, then H_2 etc.). The threshold values of Table 2 were settled upon via limited experimentation using sentences from dataset1.

Hypothesis testing was applied to both datasets. Notably, dataset1 has been used in the development of the MT systems and thus the results obtained with dataset2 should be considered the most representative ones, as they are com-

[2] www.presemt.eu

pletely unbiased and the set of sentences was unseen before the experiment and was only translated once. The number of times each hypothesis is tested for each dataset is quoted in Table 3, for both the standard (denoted as "stand") and the enriched resources ("enrich").

Parameter name	hypothesis	Exper.value
thres_h1	(H_1)	-4.50
thres_h2	(H_2)	-4.00
thres_h4	(H_4)	1.50
thres_h5	(H_5)	1.50
thres_h6	(H_6)	-5.50

Table 2. Parameter values for experiments

	Hypothesis activations per experiment			
	dataset 1		dataset 2	
Resource	stand.	enrich.	stand.	enrich
H_1	6	6	13	10
H_2	1	1	0	0
H_3	2	3	3	3
H_4	7	8	9	8
H_5	68	68	62	68
H_6	32	32	32	44

Table 3. Tested hypotheses per dataset

Since the first four hypotheses are only activated a few times each, when reporting the results, the applications of hypotheses H1 to H4 are grouped together. As hypotheses 5 and 6 are tested more frequently, the application of each one of them is reported separately.

Number of sentences	200	**Resources**	stand.
Reference translations	1	**Language pair**	EL–EN

MT config.	**Metrics**			
	BLEU	**NIST**	**Meteor**	**TER**
Baseline	0.3462	6.974	**0.3947**	51.05
H_1 **to** H_4	0.3479	6.985	0.3941	50.84
H_1 **to** H_5	0.3503	7.006	0.3944	50.80
H_1 **to** H_6	**0.3517**	**7.049**	0.3935	**50.42**

Table 4. Metric scores for dataset1, using the standard language resources, for the baseline system and for different hypotheses.

In Table 4, the results are depicted for the four MT objective evaluation metrics, when using dataset 1. For each metric, the configuration giving the highest score is depicted in boldface. As can be seen, the best BLEU score is obtained when checking all 6 hypotheses, and the same applies to NIST and TER. On the contrary, for Meteor the best result is obtained without resorting to the n-gram model information. Still the difference in Meteor scores is minor (less than 0.3%). The improvements in BLEU, NIST and TER are respectively +1.6%, +1.0% and -1.2% over the baseline, when using all 6 hypotheses. Furthermore, as the number of hypotheses to be tested increases, the performance for all three metrics is improved.

Number of sentences	200	**Resources**	enrich.
Reference translations	1	**Language pair**	EL–EN

MT config.	**Metrics**			
	BLEU	**NIST**	**Meteor**	**TER**
Baseline	0.3518	7.046	**0.3997**	50.14
H_1 **to** H_4	0.3518	7.054	0.3990	50.00
H_1 **to** H_5	0.3541	7.094	0.3995	49.72
H_1 **to** H_6	**0.3551**	**7.135**	0.3984	**49.37**

Table 5. Metric scores for dataset1, using enriched language resources, for different systems.

In Table 5, the same experiment is repeated using an enriched set of lexical resources including a bilingual lexicon with higher coverage. Notably, on a case-by-case comparison, the scores in Table 5 are higher than those of Table 4, confirming the benefits of using enriched lexical resources. Focusing on Table 5, and comparing the MT configurations without and with hypothesis testing, the results obtained are qualitatively similar to those of Table 4. Again, the best scores for Meteor are obtained when no hypotheses are tested. On the other hand, for the other metrics the n-gram modeling coupled with hypothesis testing results in an improvement to the scores obtained. The improvements obtained amount to approximately 1.0% for each one of BLEU, NIST and TER, over the baseline system scores indicating a measurable improvement.

In Tables 6 and 7, the respective experiments are reported, using dataset 2 instead of dataset 1, with (i) standard and (ii) enriched lexical resources. With standard resources (Table 6), consistent improvements are achieved as more hypotheses are activated, for both BLEU and NIST. In the case of Meteor, the best performance is obtained when no hypotheses are activated, but once again the Meteor score varies minimally (by less than 0.2%). On the contrary, the improvement obtained by activating hypothesis-checking is equal to 3.0% (BLEU), 1.4% (NIST) and 1.2% (TER). As can be seen, the improvement for previously unused dataset2 is proportionally larger than for dataset1.

Number of sentences	200	Resources	stand.
Reference translations	1	Language pair	EL–EN

MT config.	Metrics			
	BLEU	NIST	Meteor	
Baseline	0.2747	6.193	**0.3406**	Baseline
H_1 to H_4	0.2775	6.217	0.3403	H_1 to H_4
H_1 to H_5	0.2815	6.246	0.3400	H_1 to H_5
H_1 to H_6	**0.2837**	**6.280**	0.3401	H_1 to H_6

Table 6. Metric scores for dataset2, using standard language resources, for different systems.

Number of sentences	200	Resources	enrich.
Reference translations	1	Language pair	EL–EN

MT config.	Metrics			
	BLEU	NIST	Meteor	TER
Baseline	0.3008	6.541	0.3784	55.21
H_1 to H_4	0.3059	6.569	0.3790	54.96
H_1 to H_5	**0.3105**	6.593	**0.3791**	54.75
H_1 to H_6	0.3096	**6.643**	0.3779	**54.64**

Table 7. Metric scores for dataset2, using enriched language resources, for different systems.

Using the enriched resources, as indicated in Table 7, the best results for BLEU and Meteor are obtained with hypotheses 1 to 5, while for NIST and TER the best results are obtained when all six hypotheses are tested. In the case of Meteor any improvement is marginal (of the order of 0.2%). The improvements of the other metrics are more substantial, being 3.3% for BLEU, 1.6% for NIST and 1.0% for TER.

A statistical analysis has been undertaken to determine whether the additional n-gram modelling improves significantly the translation scores. More specifically, paired t-tests were carried out to determine whether the difference in translation accuracy was statistically significant, comparing the MT accuracy obtained with all six hypotheses versus the baseline system. Two populations were formed by scoring independently each translated sentence with each one of the NIST, BLEU and TER metrics, for dataset2. It was found that when using the standard resources (cf. Table 6), the translations were scored by TER to be significantly better when using the 6 hypotheses, in comparison to the baseline system, while for BLEU and NIST the translations for the 2 systems were equivalent (at a 0.05 confidence level). When using the enriched resources, no statistically significant difference was detected for any metric at a 0.05 confidence level, but significant differences were detected for all 3 metrics at a 0.10 confidence level (cf. Table 7).

5 Discussion

According to the experimental results, the addition of a new model in the hybrid MT system has contributed to an improved translation quality. These improvements have been achieved using a limited experimentation time and only a few hypotheses on what is an extensively developed language pair, for the proposed MT methodology. It is likely that as the suite of hypotheses is increased, larger improvements in objective metrics can be obtained.

When applying the hypotheses, the initial system translation is available both at token-level and at lemma-level. Out of the 6 hypotheses tested here, 5 involve token-based information and only one involves lemmas. If additional hypotheses are added operating on lemmas, a further improvement is expected.

Notably, the new n-gram modelling requires no collection or annotation of additional resources. The use of an established software package (SRILM) for assembling an n-gram database, via which hypotheses are rejected or confirmed, results in a straightforward implementation. In addition, multiple models can be effectively combined to improve translation accuracy by investigating different language aspects.

An interesting point is that the n-gram models created are factored (i.e. including information at both lemma and token level). Thus, different types of queries may be supported, to improve translation quality.

6 Future work

The experiments reported here have shown that improvements can be achieved, without specifying in detail the templates searched for, but allowing for more general formulations.

One aspect which should be addressed in future work concerns evaluation. Currently, this is limited to objective metrics. Still it is well-worth investigating the extent to which translation improvement is reflected by subjective metrics, which are the preferred instrument for quality evaluation (Callison-Burch at al., 2011).

In addition, it is possible to achieve further improvements if the hypothesis templates are made more detailed, by supplementing the lexical information by detailed PoS information.

Tests performed so far have used empirically-set parameter values for the hypotheses. It is possible to adopt a systematic methodology such as MERT or genetic algorithms to optimise the actual values of the hypotheses parameters.

Another observation concerns the manner in which the two distinct language models are applied. In the present article, n-grams are used to correct a translation already established via the phrase indexed model, having a second-level, error-checking role. It is possible, however, to revise the mode of application of the language models, so that instead of a sequential application, the two model families are consulted at the same time. This leads to an MT system that exploits the information from multiple models concurrently, and is the focus of future research.

Acknowledgements

The research leading to these results has received funding from the POLYTROPON project (KRIPIS-GSRT, MIS: 448306).

References

Chris Callison-Burch, Philip Koehn, Christof Monz, and Omar F. Zaidan. 2011. Findings of the 2011Workshop on Statistical Machine Translation. *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 22–64, Edinburgh, Scotland, UK, July 30–31, 2011.

Jaime Carbonell, Steve Klein, David Miller, Michael Steinbaum, Tomer Grassiany, and Jochen Frey. 2006. Context-Based Machine Translation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 19-28.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. *EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, Scotland, pp. 85-91.

Yannis Dologlou, Stella Markantonatou, Olga Yannoutsou, Soula Fourla, and Nikos Ioannou. 2003. Using Monolingual Corpora for Statistical Machine Translation: The METIS System. *Proceedings of the EAMT-CLAW'03 Workshop*, Dublin, Ireland, 15-17 May, pp. 61-68.

David Gale and Lloyd S. Shapley. 1962. College Admissions and the Stability of Marriage. *American Mathematical Monthly*, *Vol. 69*, pp. 9-14.

John Hutchins. 2005. Example-Based Machine Translation: a Review and Commentary. *Machine Translation*, Vol. 19, pp.197-211.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch and David Yarowsky. 2012. Toward Statistical Machine Translation without Parallel Corpora. *Proceedings of EACL2012*, Avignon, France, 23-25 April, pp. 130-140.

Philip Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Philipp Koehn and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, Vol.9, pp.9-16.

Philipp Koehn, Franz Josef Och, and Daniel Marcu, Statistical Phrase-Based Translation, *Proceedings of HLT/NAACL-2003 Conference*, Vol.1, pp.48-54.

Philip 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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL-2007 Demo & Posters Sessions,* Prague, June 2007, pp. 177-180.

Philipp Koehn, and Hieu Hoang. 2007. Factored Translation Models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* Prague, Czech Republic, pp. 868-876.

Stella Markantonatou, Sokratis Sofianopoulos, Olga Yannoutsou, and Marina Vassiliou. 2009. Hybrid Machine Translation for Low- and Middle- Density Languages. *Language Engineering for Lesser-Studied Languages*, S. Nirenburg (ed.), pp.243-274. IOS Press. ISBN: 978-1-58603-954-7

NIST 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrences Statistics.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering Foreign Language by Combining Language Models and Context Vectors. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea, Vol.1, pp.156-164.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA, pp. 311-318.

Temple F. Smith, and Michael S. Waterman. 1981. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, Vol. 147, pp. 195-197.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.

Sokratis Sofianopoulos, Marina Vassiliou, and George Tambouratzis. 2012. Implementing a language-independent MT methodology. *In Proceedings of the First Workshop on Multilingual Modeling*, held within the ACL-2012 Conference, Jeju, Republic of Korea, 13 July, pp.1-10.

George Tambouratzis, Sokratis Sofianopoulos, and Marina Vassiliou (2013) Language-independent hybrid MT with PRESEMT. *In Proceedings of HYTRA-2013 Workshop*, held within the ACL-2013 Conference, Sofia, Bulgaria, 8 August, pp. 123-130.

Vladimir I. Levenshtein (1966): Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, Vol. 10, pp. 707–710.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash (2011) SRILM at Sixteen: Update and Outlook. *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, December 2011.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu (2012) Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. *Proceedings of ACL2012*, Jeju, Republic of Korea, pp. 459-468.

Dekai Wu (2005) MT model space: Statistical versus compositional versus example-based machine translation. *Machine Translation*, Vol. 19, pp. 213-227.