Proceedings of the 16th EAMT Conference, 28-30 May 2012, Trento, Italy

Cascaded Phrase-Based Statistical Machine Translation Systems

Dan Tufiş
Research Institute for Artificial Intelligence,
Romanian Academy
Bucharest, Romania
tufis@racai.ro

Ştefan Daniel Dumitrescu
Research Institute for Artificial Intelligence,
Romanian Academy
Bucharest, Romania
sdumitrescu@racai.ro

Abstract

Statistical-based methods are the prevalent approaches for implementing machine translation systems today. However the resulted translations are usually flawed to some degree. We assume that a statistical baseline system can be re-used to automatically learn how to (partially) correct translation errors, i.e. to turn a "broken" target translation into a better one. By training and testing on initial bilingual data, we constructed a system S1 which was used to translate the source language part of the training corpus. The new translated corpus and its reference translation are used to train and test another similar system S2. Without any additional data, the chain S1+S2 shows a sensible quality increase against S1 in terms of BLEU scores, for both translation directions (English to Romanian and Romanian to English).

1 Introduction

The paper presents a cascaded phrase based translation system that obtains improved translation scores using no additional data compared to the standard single-step translation system.

The first challenge of our research was to obtain the best standard translation system possible. We experimented with different factored models that include surface form, lemmas and different part of speech tag sets in various combinations to confirm the assumption that translation accuracy is improved over a surface form only baseline model.

The second objective of our work was to validate our intuition that a statistical baseline system can be re-used (cascaded) to automatically learn how to (partially) correct its own translation errors, i.e. to turn an initially "broken" translation into a better one.

The phrase-based translation approach has overcome several drawbacks of the word-based translation methods and proved to significantly improve the quality of translated output. The morphology of a highly inflected language permits a flexible word order, thus shifting the focus from long-range reordering to the correct selection of a morphological variant.

Morphologically rich languages have a large number of surface forms in the lexicon to compensate for a flexible word order.

Both Transfer and Interlingua MT employ a generation step to produce the surface form from a given context and a lemma of the word. In order to allow the same type of flexibility in using the morpho-syntactic information in translation, factored translation models (Koehn and Hoang, 2007) provide the possibility to integrate the linguistic information into the phrase-based translation model.

Most of the statistical machine translation (SMT) approaches that have a morphologically rich language as target employ factored translation models. Our approach is similar to several other factored machine translation experiments such as adding the morphological features as factors (Avramidis and Koehn, 2008). Our results confirm findings of other researchers, namely that when very large parallel corpora are available, minimal pre-processing is sufficient to get better results than the baseline (raw data); however, when only a limited amount of training data is available, better results are achieved with part-of-speech tags and complex morphological analysis (Habash and Sadat, 2006).

Romanian is a morphologically rich language which needs more than 1200 lexical tags in order to be compliant with the Multext-East lexical specifications (Erjavec and Monachini, 1997).

Czech and Slovene require more than 2000 such morpho-lexical descriptors (MSDs). These descriptors encode detailed linguistic information (gender, case, modality, tense etc.) which can be extremely useful for an accurate translation based on factored models. The set of MSDs can be reduced without information loss by exploiting the redundancy between various feature-value combinations in these descriptors. Yet, the resulting tagsets are too large and thus the data-sparseness hampers the reliability of automatic assignment of MSDs to arbitrary new texts.

Tiered tagging (Tufiș, 1999) is a two-stage technique addressing the issue of training data sparseness. It uses an automatically induced intermediary tag-set, named CTAG tagset, of a smaller size on the basis of which a common POS tagging technique can be used. In a second phase, it replaces the tags from the small tag-set with tags from the fully-specified morpho-syntactic tag-set (MSD tag-set) also taking into consideration the context. The second phase of tiered tagging relies on a lexicon and a set of hand-written rules. The original idea of tiered tagging has been extended in (Ceaușu, 2006), so that the second phase is replaced with a maximum entropy-based MSD recovery. In this approach, the rules for CTAG to MSD conversion are automatically learnt from the corpus. Therefore, even the CTAG labels assigned to unknown words can be converted into MSD tags. If an MSD-lexicon is available, replacing the CTAG label for the known words by the appropriate MSD tags is almost 100% accurate.

2 System overview

Factored translation models extend the phrase-based translation by taking into account not only the surface form of the phrase, but also additional information like the dictionary form (lemma), the part-of-speech tag or the morpho-syntactic specification. It also provides, on the target side, the possibility to add a generation step. All these new features accommodate well in the log-linear model employed by many decoders:

$$P(e|f) = exp \sum_{i=1}^{n} \lambda_i h_i(e,f) \quad (1)$$

where $h_i(e,f)$ is a function associated with the pair e, f and λ_i is the weight of the function.

To improve the translation into morphologically-rich languages, the multitude of options provided by the factored translation can help validate the following assumptions:

a) Aligning and translating *lemma* could significantly reduce the number of translation equivalency classes, especially for languages with rich morphology;

b) *Part of speech affinities*. In general, the translated words tend to preserve their part of speech and when this is not the case, the part-of-speech chosen is not random;

c) The *re-ordering* of the target sentence words can be improved if language models over POS or MSD tags are used.

In order to test the improvement of the factored model over the phrase-based approach, we built strong baseline systems for the RO-EN language pair (Ceaușu and Tufiș, 2011).

The intuition that motivated our experiments is that the same methodology used in translating from language A into language B could be applied for (partially) correcting the initial translation errors. We wanted to validate this idea without recourse to additional resources. To this end, we built a two – layered cascaded translation system.

The first step was to create the best possible direct translation system S1 for A→B. For this we started from a parallel corpus: {C_A,C_B}. Using this corpus we trained a factored phrased-based translation model. Having the A→B system obtained (Ceaușu and Tufiș, 2011), we prepared for the second system S2 by translating the entire training corpus C_A into language B, obtaining $T_{S1}(C_A)$. Using the new parallel corpus {$T_{S1}(C_A)$,C_B} we trained the second system S2.

At this point we chained the two systems together: we give an input text I_A (in language A), the first system translates I_A to $T_{S1}(I_A)$ which is the input for the second system. Thus, the chained system receiving the input I_A produces the output O_B:$T_{S2}(T_{S1}(I_A))$.

We further present the steps taken to build this cascaded system and compare the translation performance against the direct, S1 one-step system.

3 Data Preparation

The corpus used to train any SMT system has the biggest influence on translation quality, so special attention is given to its preparation. For the purposes of this paper we used the bilingual parallel corpus (Romanian-English) that had been developed during the ACCURAT FP-7 research project. We chose this resource because it is a reasonably large parallel corpus between a highly inflectional language (Romanian) and a less inflectional reference language (English).

The content of the corpus is drawn from several other corpora:

1) DGT-TM[1], law and juridical domain, approx. 650,000 sentences;

2) EMEA (Tiedemann, 2009), medical corpus, approx. 994,000 sentences;

3) Romanian-English part of the multilingual thesaurus Eurovoc[2], (1-5 words), approx. 6,500 bilingual terms, treated as short sentences;

4) PHP[3], translation of the PHP software manual, approx. 30,000 sentences;

5) KDE[4], translation of the Linux KDE interface, approx. 114,000 sentences;

6) SETIMES [5] , news corpus, approx. 170,000 sentences.

In total, the source Romanian – English corpus has over 1,950,000 sentences. However, the corpus needed to be cleaned and annotated. This was performed in three steps:

1) Step 1 – Initial corpus cleaning – We created a cleaning application that removes duplicate lines (ex: the PHP corpus contains many identical lines), lines that contain only/mostly numbers (such as lines that consist only of telephone numbers), lines that contain no Latin characters, lines that contain less than 3 characters and other similar heuristics. Additionally, there are three specific types of text distortions occurring in Romanian texts: (i) missing diacritical characters, (ii) different encoding codes for the same diacritical characters, and (iii) different orthographic systems. When ignored, they have a negative impact on the quality of translation and language models and, thus, on the translation results. For details on the process of diacritics restoration, see (Tufiș and Ceaușu, 2008).

2) Step 2 – Corpus annotation – The parallel corpus was annotated using our NLP tools (Tufiș et al., 2008) that tokenize, lemmatize and tag the input text. The tagger does its job both in terms of CTAG and MSD tagsets. This annotation was performed for both Romanian and English sides of the corpus. The annotation has the Moses (Koehn et al., 2007) input file structure. For example, the sentence: “Store in the original package.” has been annotated as shown in Table 1, one token per line followed by three additional fields, separated by “|”:

<table>
<tr><th><i>English</i></th><th><i>Romanian</i></th></tr>
<tr><td><b>Store</b>|store^Nc|NN|Ncns</td><td><b>A</b>|avea^Va|VA3S|Va--3s</td></tr>
<tr><td><b>in</b>|in^Sp|PREP|Sp</td><td><b>se</b>|sine^Px|PXA|Px3--a--------w</td></tr>
<tr><td><b>the</b>| the^Dd| DM|Dd</td><td><b>pastra</b>|pastra^Vm|V3|Vmii3s</td></tr>
<tr><td><b>original</b>|original^Af|</td><td><b>în</b>|în^Sp|S|Spsa</td></tr>
<tr><td><b>package</b>|package^Nc|NN|Ncns</td><td><b>ambalajul</b>|ambalaj^Nc|NSRY|Ncmsry</td></tr>
<tr><td></td><td><b>original</b>|original^Af|ASN| Afpms-n</td></tr>
</table>

Table 1. EN-RO annotated sentence pair

0 – surface form – the token itself;
1 – lemma of the token, trailed (^) with the grammar category;
2 – CTAG – tag from the reduced tagset;
3 – MSD – Morpho-Syntactic Annotation tag.

3) Step 3 – Final cleaning – The last step involved using the Moses cleaner, a Perl script that ensured that the corpus did not contain illegal characters, spaces, etc. and that the two corpus sides (Romanian – English) had an equal number of sentences.

After these cleaning steps the RO-EN corpus was reduced to around 1,250,000 sentences. Finally, the corpus was randomized and 1200 sentence-pairs (T_{RO}-T_{EN}) were extracted that represent the RO-EN test files.

4 Translation experiments

5.1 First layer translation system (S1)

The first step was to decide on a model for the direct Romanian ↔ English translation. Several models have been proposed and tested. Using the Moses SMT software, we have created the following models (we have experimented with several more models, but kept here only the top performers for reference):

Model #	*Details*
#1	t0-0 m0
#2	t1-1 g1-0 m0
#3	t1-1 g1-3 t3-3 g1,3-0 , m0m3
#4	t1-1 g1-3 t3-3 g1,3-0 , m0m3 r0
#5	t1-1 g1-3 t3-3 g1,3-0 , m0m3 r3

Table 2. Models description for the first layer

Notation: t = translation step, g = generation step, m = language model, r = reordering model. The first model (#1) simply translates surface forms in language A to surface forms in language

[1] http://langtech.jrc.it/DGT-TM.html
[2] http://eurovoc.europa.eu/
[3] http://www.php.net/
[4] http://docs.kde.org/
[5] http://www.setimes.com/

B (t0-0). The second model (#2) first translates lemmas in language A to lemmas in language B (t1-1) and then employs a generation step to generate surface forms in language B from lemmas in language B (g1-0). The third, fourth and fifth models (#3, #4, #5) follow a more complex path. They first start with a lemma-lemma translation (t1-1), followed by a lemma to MSD generation in language B (g1-3), a translation of MSDs in language A to MSDs in language B (t3-3) and finally generating surface forms from the previously translated lemmas and MSDs in language B (g1,3-0). They use two language models. While models #1 and #2 use just a surface language model, models #3, #4 and #5 additionally use a MSD language model. The difference between models #3, #4 and #5 is that model #4 uses a reordering model based on surface forms while model #5 uses reordering based on MSDs. Table 3 presents the BLEU scores (Papineni et al., 2002) obtained testing the five proposed models.

For the Romanian → English direction, model #3 was the best performing of the five, with a BLEU score of 57.01. For the English → Romanian direction, scores were a bit lower, model #2 having the highest 53.94 BLEU points.

Interestingly, the large size of the corpus shows its power, bringing the score of the unfactored model #1 very close to the factored models.

The next step was to estimate the translation time of the corpus. This was necessary because of the size of the training corpus: approx. 1.25 million sentences. Moses offers two different translation options: the default translation search and the cube pruning search algorithm. There are two adjustable parameters: the stack size and beam search. These parameters have been manually specified to obtain insights about their influence on translation speed and quality. We present only model #3 for the RO→EN direction.

The translation time includes language model and translation/generation tables loading time. The test machine is a dedicated 16 core (8 physical + 8 virtual, running at 2.6GHz), 12 GM RAM server.

RO → EN		*EN → RO*	
Model #	*BLEU*	*Model #*	*BLEU*
#1	56.31	#1	52.43
#2	56.49	**#2**	**53.94**
#3	**57.01**	#3	49.97
#4	56.79	#4	49.12
#5	56.89	#5	48.70

Table 3. S1: Model scores

Stack Size Param.	*Beam Search Param.*	*Translation Time (s)*	*BLEU Score*
(default)	(default)	3074	**57.01**
100	(default)	1611	56.69
50	(default)	831	56.05
20	(default)	391	54.97
15	(default)	307	54.36
10	(default)	229	53.16
5	(default)	144	51.35
(default)	100	83	39.17
(default)	10	83	43.29
(default)	2	87	47.17
(default)	1	93	49.63
(default)	0.5	151	51.80
(default)	0.1	169	55.84
100	1	106	49.63
Cube pruning algorithm with stack size 2000		**167**	56.29

Table 4. S1: Parameter variation, translation time and BLEU scores.

Table 4 shows measurements for the translation times and BLEU scores (RO→EN direction) of the test files (1200 sentences), for different settings of the Stack Size and Beam Search.

Even though the best performing translation was achieved using the default parameters (BLEU score: 57.01), due to the very long translation time, we found that the best compromise was to use the cube pruning algorithm with the stack size 2000 that obtains a marginally lower BLEU score of 56.29. When using the cube pruning algorithm, we found that, for our test set, increasing the stack size to more than 2000 does not generate any noticeable score improvements.

Based on these results, we have used the two best performing models (model #3 for the RO→EN direction and model #2 for the EN→RO direction) with the cube pruning search algorithm to translate both languages of the parallel corpus $\{C_{RO}, C_{EN}\}$. We obtained two new corpora: for the RO→EN direction we obtained the $\{T_{S1}(C_{RO}),C_{EN}\}$ corpus, and for the EN→RO direction we obtained the $\{C_{RO},T_{S1}(C_{EN})\}$ corpus.

After the translation, the final phase of this step was to process the two newly obtained corpora. Using the same NLP tool we used to annotate the original corpus we annotated the translated corpora with lemma, CTAGs and MSDs. Finally, the annotated corpora were cleaned again, but using only step 3 (the Moses cleaning script) of the cleaning process described in section 3. The cleaning yielded for the RO→EN direction a corpus of around 1,110,000 sentences (losing in this second cleaning process about

140,000 sentences - around 11% - from the initial 1,250,000), while for the EN→RO direction the corpus lost almost 240,000 sentences resulting in a corpus of 1,010,000 sentences.

5.2 Second layer translation system (S2)

For this step, using the intermediary corpus, we trained 9 models to see which one would perform best. Table 5 shows the models chosen and table 6 shows the translation and BLEU scores using the cube pruning and default translation algorithms. The same models were used for both translation directions.

Model	*Details*
#1	t0-0 m0
#2	t1-1 g1-0 m0
#3	t1-1 g1-2 t2-2 g1,2-0 m0,m2
#4	t1-1 g1-3 t3-3 g1,3-0 m0,m3
#5	t1-1 g1-3 t3-3 g1,3-0 m0,m3 r3
#6	t1-1 g1-2 t2-2 g2-3 t3-3 g1,3-0 m0,m2,m3
#7	t0,1-0,1 m0
#8	t0,1,2-0,1,2 m0,m2
#9	t1,2-t1,2 m0,m2

Table 5. S2: Models description

Translating was performed with both default parameters and using the cube pruning search with stack size 2000. The reordering model is the Moses default, with the only difference that in model 5 we have used MSDs as the reordering factor.

For testing S2 we used the same test files as for S1, but translated with the best S1 models: the model #3 for RO→EN direction and the model #2 for the EN→RO direction (see Table 3). The reference translations for the two directions were T_{EN} and T_{RO} respectively (1200 sentences each).

For the RO→EN direction the BLEU translation score of the S1+S2 system has been improved from the best S1 model (57.01) to a new BLEU score of 60.90.

The fact that S2 translation based on model #7 (surface form & lemma to surface form & lemma using only the surface language model) was the fastest and most accurate is not surprising: we "translated" from partly broken English into presumably better English.

Generation steps were not necessary and the information on the lemma eliminated some candidates from the search space.

Interestingly, the translation time the using default Moses parameters is very close to the cube pruning search (because the chosen model has just phrase translation and no generation component), but yields approximately 0.14 BLEU point increase.

Model #	*Transl. time (s) vith cube pruning*	*BLEU with cube pruning*	*Transl. time (s) with default params.*	*BLEU with default params.*
#1	195	60.42	257	60.65
#2	186	59.59	4745	60.12
#3	175	55.68	4129	56.12
#4	281	55.50	3994	56.18
#5	221	55.45	4104	56.20
#6	244	55.16	5016	55.98
#7	**108**	60.74	143	**60.90**
#8	144	58.50	254	58.61
#9	136	58.50	249	58.61

Table 6. RO→EN: *S2(S1(T_{RO}))*

Table 7 shows that for the EN→RO direction, the S2 system models #7 and #8 have a similar performance, increasing the BLEU score from the original 53.94 points to 54.44 (0.5 BLEU point net increase). As with the RO→EN direction, the S2 models that employ generation steps actually slightly decrease the score.

Model #	*Transl. time (s) vith cube pruning*	*BLEU with cube pruning*	*Transl. time (s) with default params.*	*BLEU with default params.*
#1	254	54.41	154	54.42
#2	1443	52.14	556	52.55
#3	1051	53.50	594	53.50
#4	543	53.59	798	53.59
#5	530	53.59	613	53.59
#6	805	53.56	997	53.56
#7	282	54.43	**167**	**54.44**
#8	417	54.41	287	**54.44**
#9	403	54.40	280	54.42

Table 7. EN →RO: *S2(S1(T_{EN}))*

6 Evaluation procedure and discussion

After the original corpus was annotated and cleaned, it was split into two separate files for each language: training set and test set. The test file T_{EN}-T_{RO} contains 1200 aligned sentences. Since the sentences were extracted from the randomized corpus after cleaning, the test files contain sentences from all genres that make up the original corpus, so they represent **in-domain** data.

In Tables 6 and 7 we showed that the cascaded factored SMT (S1+S2) performs better than the baseline system (S1) for both translation directions, in terms of BLEU scores. We were interested to see which were the most distant translations from the reference, assuming that these were bad translations. We computed for each sentence *I* the similarity scores SIM between its translations and the reference translation. These scores were computed with the same BLEU-4 function used for bitexts. Similarly to the BLEU score applied to a bitext, 100 means perfect match and 0 means complete mismatch. Thus, we obtained 1200 pairs of scores SIM_{S1}^{I} and SIM_{S1+S2}^{I}. We also compute the average similarity scores as $\frac{1}{1200}\sum_{I=1}^{1200} SIM_{S\alpha}^{I}$ where S_{α} is S1 or S1+S2. As expected, the average SIM scores make the same ranking as the BLEU scores, although they are a bit higher (ex: 61.18 for S1 and 63.58 for S1+S2 for the RO→EN direction).

We briefly comment on the results of this analysis for the Romanian-English translation direction. We manually analysed the test set translations. We identified 3 sentences with their translations having a zero SIM score for both systems. The explanation was that the reference translation was wrongly aligned to the source sentence.

S1 produced 72 perfect translations (score 100) while S1+S2 produced 105. Only 57 perfect translations were common to S1 and S1+S2, meaning that S1+S2 actually deteriorated a few of the original perfect translations. By analyzing the 15 translations that were "deteriorated" we noticed that they were identical, except that unlike S1+S2, S1 and Reference translations either had a differently capitalized letter that marginally lowered the score or had multiword units joined by underscores (e.g. *as well as* vs. *as_well_as*). This was a small bug which has been removed and which, overall, brought a 0.05 increase in the BLEU score. One of the "degraded" translation pair is given below:

RO: *după examinarea problemelor și consecințelor posibile , Uniunea Democrată Croată a Primului Ministru Ivo Sanader și aliații săi parlamentari au decis să sprijine amânarea .*

S1: *after examination problems and possible consequences , the Democratic Union of Croatian Prime_Minister Ivo Sanader and his allies lawmakers decided to support the postponement .* (score 0.1794)

S1+S2: *after examination problems and possible consequences , the Croatian Democratic Union of Prime Minister Ivo Sanader and his allies lawmakers decided to support the postponement .* (score 0.1695)

EN$_{REF}$: *after considering possible issues and consequences , Prime_Minister Ivo Sanader 's Croatian Democratic Union and its parliamentary allies decided to support a delay . "*

If one ignores the underscore issue in the S1+S2 translation, then this translation is better than the one of S1. A frequent translation difference with respect to the reference translations is illustrated by the example above: the Saxon genitive construction for noun phrases is replaced by a prepositional genitival construction (in this case the word order is closer to the Romanian word order).

The capitalization and punctuation are other sources of lower scoring against the reference. All these examples show the sensitivity of the BLEU scoring method, especially for very short sentences.

Another important variable to note is the amount of change from one layer to the other: out of all sentences, around 37% had a BLEU increase while around 20% had a BLEU decrease (but see the comment on the underscore difference), the rest 43% have not been changed in any way.

Overall, we obtain a 3.89 BLEU point increase for the RO→EN direction and a smaller 0.5 BLEU point increase for the more difficult EN→RO direction using our cascaded system.

Another interesting result was to evaluate the simple cascading systems without feature models, that is (S1=t0-0m0)+(S2=t0-0m0) and compare their performances with the direct translations and the best feature-models cascaded systems. The results are shown in Table 8.

RO →EN'→EN		*EN →RO'→RO*	
Model #	*BLEU*	*Model #*	*BLEU*
#1+#1	60.47	#1+#1	54.29
#3+#7	60.90	#2+#7	54.44

Table 8. *S2(S1(T$_{source}$))*

The increased accuracy due to various feature combinations versus the baseline system has been apparent from Tables 6 and 7 compared to the results in Table 3. Table 8 shows that the direct translations (S1 with any model) for both directions have BLEU scores lower than the cascaded system (S1+S2) even when feature models were not used (model #1+#1).

Thus, we can support the statement that the morphological features and the cascading idea are beneficial to the overall accuracy of translations (at least between Romanian and English).

S1 SIM *S2 SIM* *Difference*	*Romanian Source*	*S1 Translation*	*S2 Translation*	*English Reference*
0.397 0.492 *0.095*	bun , și-acuma să revenim la problema lui cum și de ce.	good , and now *to **revenim*** to the problem of how and why.	good , and now *let us go* to the problem of how and why.	and now let us get back to the question of how and why.
0.392 0.660 *0.268*	spune-mi ce crezi tu că-ți amintești.	tell me what *believe you* that you remember.	tell me what *you think* that you remember.	tell me what you think you remember.
0.213 0.316 *0.104*	În primul rând , pentru că mărturisirile pe care le făceau erau evident smulse și neadevărate.	firstly , because confessions *on which* they made were obviously clean and jerk and untrue.	firstly , because *the* confessions they made were obviously clean and jerk and untrue.	in the first place , because the confessions that they had made were obviously extorted and untrue.
0.447 0.376 *-0.071*	cum ar putea muri ?	how *could die* ?	how *to die* ?	how could he die ?
0.256 0.216 *-0.039*	cei trei nu făcuseră nici o mișcare.	the three not ***făcuseră*** any movement.	the three not *to make* any movement.	the three men never stirred.

Table 9. Out-of-domain text S1 / S1+S2 translation improvement / degradation examples for RO→EN

Given the corpus is almost entirely composed of juridical and medical texts, we were anxious to see how the second translation step would perform on **out-of-domain** texts.

To make things even harder, we chose a different genre: literary fiction. We extracted 1000 sentences between 3 and 40 words long from Orwell's "1984" novel. This test text is challenging because it contains many out of vocabulary words, new senses, frequent subject-elided constructions (Romanian is a pro-drop language), verbal tenses specific to literary narratives which are practically absent from the training data. Another challenge was due to the Romanian translation of Orwell's original, which is not a word-for-word translation, but a literary one.

We tested only the RO→EN direction with the following results: the first translation system (S1) obtained a score of 27.53 BLEU points (model #3), while the second system (S2) marginally improved the translation to 27.70.

Out of the 1000 sentences, 69 have had their scores properly increased and 76 slightly "decreased". However, even if the overall BLEU score increase was minimal, we observed that the translation quality has improved from a human analysis point of view. The positive and negative examples (Table 9) show that even though the changes in SIM score are minimal, the text produced by S2 corrects some of the unknown words of S1 (by synonyms or paraphrases, not matching the reference) as well as phrase structure by better word choice and word reordering (corrections missed by the BLEU/SIM scores).

Finally, we took the cascading idea one step further by repeating the entire train-translate process (step 2), obtaining $S3(S2(S1(T_{source})))$. We observed that the translation stabilized, with very few sentences being changed (around 1%), and with the changes being minor (increasing or even decreasing the BLEU score by less than ~0.05 points). We concluded that further cascading would not bring significant improvements.

7 Conclusions and future work

This article presented a simple but effective way of further improving the quality of a phrased-based statistical machine translation system, by cascading translators. We are not aware of better translation scores for the Romanian-English pair of languages. The idea of post-processing the output of a SMT system is not new but, this step was most often than not based on hand-crafted rules or other knowledge intensive methods. A similar idea was recently reported in (Ehara, 2011) but, their EIWA ensemble is based on a commercial rule-based MT (specialized in patent translation) for the first step and a MOSES-based SMT for the second phase (named statistical post-editing). There are several other methodological differences between our system and the one described in (Ehara, 2011). EIWA does not work in real time because before proper translation of a text T, the SMT post-editor is trained on a text similar to T. The similar text is constructed from a large patent parallel corpus (3,186,284 sentence pairs) by selecting for each sentence in T an average number of 127 similar sentences.

We use the same SMT system trained on different parallel data. The first system S1, trained on parallel data $\{C_A,C_B\}$ learnt to produce draft translations from L_A to L_B. The second translation system S2, trained on the "parallel" data $\{S1(C_A)\text{-}C_B\}$, learnt how to improve the draft translations. Except for the training data and the different parameter settings, the two systems are incarnations of the same basic system. Contrary to Ehara (2011), we found that setting the distortion parameter to a non-null value improves the translation quality. Translation of a new, unseen text is achieved in real time (no retraining at the translation time). While in (Ehara, 2011) improvements were reported for two language pairs (Japanese to English and Chinese to English), we showed that our approach, for the present moment, works only for one language pair (Romanian and English) but in both translation directions. We also showed that the cascaded approach improves the translation quality for both in-domain and out-of-domain texts, although not to the same degree.

As future research, we are considering extending the factored experiment with comparable parallel data. The comparable data is available through the ACCURAT project. The aim of the ACCURAT project, to be finalized in June this year, is to research methods and techniques to overcome one of the central problems of machine translation (MT) – the lack of linguistic resources for under-resourced areas of machine translation. Within this context various narrow domain adaptation techniques will be evaluated and experiments will be conducted for several other language pairs.

Acknowledgments. This work has been supported by the ACCURAT project (www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement n° 248347.

References

Avramidis E., Koehn, P. 2008. Enriching morphologically poor languages for statistical machine translation. In: *Proceedings of Association for Computational Linguistics / HLT*, pp. 763–770, Columbus, Ohio

Ceaușu Alexandru. 2006. Maximum Entropy Tiered Tagging, Janneke Huitink & Sophia Katrenko (eds), *Proceedings of the Eleventh ESSLLI Student Session*, ESSLLI 2006, pp. 173-179

Ceaușu, A., Tufiș, D. 2011. Addressing SMT Data Sparseness when Translating into Morphologically-Rich Languages. In Bernadette Sharp, Michael Zock, Michael Carl, and Arnt Lykke Jakobsen (eds.) *Human-machine interaction in translation*, Copenhagen Business School, pp. 57-68.

Ehara T. 2011. Machine translation system for patent documents combining rule-based translation and statistical postediting applied to the PatentMT Task, *Proceedings of NTCIR-9 Workshop Meeting*, December 6-9, 2011, Tokyo, Japan, pp. 623-628.

Erjavec, T., Monachini, M. (Eds.). 1997. *Specifications and Notation for Lexicon Encoding.* Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/ mte-d11f/

Habash, N., Dorr, B., Monz, C. 2006. Challenges in Building an Arabic-English GHMT System with SMT Components. In *Proceedings of AMTA'06*, Cambridge, MA, USA.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, demonstration session, Prague.

Koehn, P., Hoang, H. 2007. Factored Translation Models. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 868–876, Prague.

Papineni, K., Roukos, S., Ward, T., Zhu W.J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 311-318.

Tiedemann, J. 2009. News from OPUS - *A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. In: N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pp. 237-248.

Tufiș, D. 1999. Tiered Tagging and Combined Classifiers. In: F. Jelinek, E. Nth (eds) *Text, Speech and Dialogue* LNCS vol. 1692, pp. 28-33 Springer-Verlag Berlin Heidelberg.

Tufiș, D., Ceaușu, A. 2008. DIAC+: A Professional Diacritics Recovering System, in *Proceedings of LREC 2008*, May 26 - June 1, Marrakech, Morocco. ELRA - European Language Resources Association.

Tufiș, D., Ion, R., Ceaușu, A., Ștefănescu, D. 2008. RACAI's Linguistic Web Services, in *Proceedings of LREC 2008*, May 26 - June 1, Marrakech, Morocco. ELRA - European Language Resources Association.