Can we relearn an RBMT system?

Loïc Dugast (1,2) **Jean Senellart** (1) **Philipp Koehn** (2)
dugast@systran.fr senellart@systran.fr pkoehn@inf.ed.ac.uk

(1) SYSTRAN S.A.
La Grande Arche
1, Parvis de la Défense
92044 Paris
La Défense Cedex
France

(2) School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
United Kingdom

Abstract

This paper describes SYSTRAN submissions for the shared task of the third Workshop on Statistical Machine Translation at ACL. Our main contribution consists in a French-English statistical model trained without the use of any human-translated parallel corpus. In substitution, we translated a monolingual corpus with SYSTRAN rule-based translation engine to produce the parallel corpus. The results are provided herein, along with a measure of error analysis.

1 Introduction

Current machine translation systems follow two different lines of research: (1) manually written rules associated with bilingual dictionaries (rule-based systems), (2) a statistical framework (statistical machine translation) based on large amount of monolingual and parallel corpora. The first line uses linguistically generalized information based on what humans understand from what happens in a given language (source and target) and what happens in the translation process. The translation process is *building* a translation from a given source sentence based on this knowledge. The second line exploits implicit information present in already translated corpora and more generally any text production in the target language to automatically *find* the most likely translation for a given source sentence. This approach has proven to be competitive with the rule-based approach when provided with enough resources on a specific domain. Though based on fundamentally different paradigms and exploiting different types of information, these two research lines are not in opposition and may be combined to produce improved results. For instance, serial combination of the two approaches has produced very good results in WMT07 (Simard, 2007), (Dugast, 2007) and NIST07 (Ueffing, 2008). (Schwenk et al., 2008) also combines both approaches and resources to build a better system.

The SYSTRAN's R&D team actually works to merge these two approaches, drawing benefit from their respective strengths. Initially, the SYSTRAN system was a pure rule-based system that in recent years began integrating statistical features and corpus-based model (Senellart, 2006). It must be noted that, for sake of simplification of the experiment and its interpretation, the base system mentioned in this paper is a purely rule-based version. In the framework of this research effort, various exploratory experiments are being run which aim both at finding efficient combination setups and at discriminating strengths and weaknesses of rule-based and statistical systems.

We had performed a first analysis on a statistical post-editing system (Dugast, 2007). The system submitted for Czech-English follows this setup. We present also here an original French-English statistical model which doesn't make use of the target side of the parallel data to train its phrase-table, but rather uses the rule-based translation of the source side. We call this system "SYSTRAN Relearnt" because, as far as the translation model is concerned, this system is a statistical model of the rule-based engine. In addition to the submitted system which only makes use of the Europarl monolingual data, we present additional results

Proceedings of the Third Workshop on Statistical Machine Translation, pages 175–178,
Columbus, June 2008.

using unrelated monolingual data in the news domain. Though human evaluation of these systems will provide additional insight, we try here to start analyzing the specificities of those systems.

2 Training without any human reference translation

If the need in terms of monolingual corpus to build language models can most of the time be fulfilled without much problem, the reliance of statistical models on parallel corpora is much more problematic. Work on domain adaptation for statistical machine translation (Koehn and Schroeder, 2007) tries to bring solutions to this issue. Statistical Post-Editing may well be another way to perform efficient domain-adaptation, but still requires parallel corpora. We try here to open a new path. Our submitted system for French-English on the Europarl task is a phrase based system, whose phrase table was trained on the rule based translation of the French Europarl corpus. The French side of the Europarl parallel corpus was translated with the baseline rule-based translation engine to produce the target side of the training data. However, the language model was trained on the real English Europarl data provided for the shared task. Training was otherwise performed according to baseline recommendations.

Corpus	Size (sentences)	Size (words)
Parallel FR-EN	0.94 M	21 M
Monolingual EN (LM)	1.4 M	38 M

Table 1: Corpus sizes for the submitted Europarl-domain translation

An additional (non-submitted) system was trained using two monolingual news corpora of approximately a million sentences. The French corpus was built from a leading French newspaper, the English from a leading American newspaper, both of the same year (1995). In the previous model, the English corpus used to train the language model actually contained the reference translations of the source corpus. This is not the case here. As for the previous model, the French corpus was translated by the rule-based system to produce the parallel training data, while the English corpus was used to train a language model,. This same language model is used in both statistical models: a *relearnt* system and a baseline phrase-based model whose phrase table was learnt from the Europarl parallel data. Both trainings followed the baseline recommendations of the shared task.

Corpus	Size (sentences)	Size (words)
Parallel FR-EN (Europarl v3)	0.94M	21M
Monolingual FR (Le Monde 1995)	0.96M	18M
Monolingual EN (NYT 1995)	3.8M	19M

Table 2: Corpus sizes for the additional model, trained on news domain

3 Results for the SYSTRAN-relearnt systems

We provide here results on evaluation metrics, an initial error analysis and results on the additional *relearnt* model.
Table 3 provides metrics results for four different systems : purely rule based, purely statistical, and the *relearnt* systems: Relearnt-0 is a plain statistical model of systran, while Relearnt uses a real English language model and is tuned on real English.

Model	BLEU(tuning, dev2006)	BLEU (test, devtest2006)
Baseline SYSTRAN	n.a.	21.27
Relearnt-0, with SYSTRAN English LM, tuned on SYSTRAN English	20.54	20.92
Relearnt	26.74	26.57
Baseline Moses	29.98	29.86

Table 3: Results of systems on Europarl task, trained (when relevant) on Europarl-only data

The score of the Relearnt-0 model is slightly lower than the rule-based original (absence of morphological analysis and some non-local rules which failed to be modelled may explain this). The use of a real English language model and tuning set gives a more than 5 BLEU points improvement, which is only 3 BLEU points below the Moses baseline, which uses the Europarl phrase table.

Comparing these three systems may help us discriminate between the statistical nature of a translation system and the fact it was trained on the relevant domain. For this purpose, we defined 11 error types and counted occurrences for 100 random-picked sentences of the devtest2006 test corpus for the three following systems : a baseline phrase-based system, a SYSTRAN relearnt phrase-based system and the baseline SYSTRAN rule-based system. Results are displayed in tables 5.a and 5.b.

MC	Missing Content
MO	Missing Other
TCL	Translation Choice (content, lemma)
TCI	Translation Choice (content, inflection)
TCO	Translation Choice (other)
EWC	Extra Word Content
EWO	Extra Word Other
UW	Unknown word
WOS	Word Order, short
WOL	Word Order, long (distance>=3 words)
PNC	Punctuation

Table 4 : Short definition of error types

System	MC	MO	TCL	TCI	TCO
SYSTRAN	**0.02**	**0.2**	1.11	**0.14**	0.48
Relearnt	0.22	0.39	0.77	0.22	0.38
Moses	0.35	0.46	**0.63**	0.27	**0.25**

Table 5.a : Average number of errors/sentence

System	EWC	EW0	UW	WOS	WOL	PNC
SYSTRAN	**0**	0.72	**0.06**	0.41	**0.02**	**0**
Relearnt	0.05	**0.35**	0.09	0.41	0.05	**0**
Moses	0.17	0.4	0.12	**0.3**	0.08	0.02

Table 5.b : Average number of errors/sentence

Such results lead us to make the following comments, regarding the various error types:

- Missing words

This type of error seems to be specific to statistical systems (counts are close between *relearnt* and baseline Moses) . Although we do not have evidence for that, we guess that it is especially impairing adequacy when content words are concerned.

- Extra words

Obviously, the rule-based output produces many useless functional words (determiners, prepositions…) while statistical systems do not have this problem that much. However, they may also produce extra content words..

- Unknown words

Few words are out of the rule-based dictionaries' vocabulary. Morphological analysis may explain at least part of this.

- Translation choice

Translation choice is the major strength of the statistical model. Note that the *relearnt* system gains a great deal of the difference between Systran and Moses in this category. We would expect the remaining difference to require more translation choices (which may be learnt from a parallel corpus). Inflection errors remain low for the rule-based system only, thanks to its morphological module.

- Word Order

The language model couldn't lower the number of short-distance word-order errors (no difference between SYSTRAN and SYSTRAN relearnt). Long-distance word order is, as expected, better for the rule-based output, though French-English is not known to be especially sensitive to this issue.

Additionally, table 6 shows the results of the *relearnt* system we trained using only monolingual corpus. It performed better than both the europarl-trained phrase-based model and the baseline rule-based engine. Table 7 shows the three different translations of a same example French sentence.

Model	BLEU (tuning, nc-dev2007)	BLEU (test, nctest2007)
SYSTRAN	n.a.	21.32
Relearnt	22.8	23.15
Baseline Moses	22.7	22.19

Table 6 : Results of systems on News task

SOURCE	Ces politiques sont considérées comme un moyen d'offrir des réparations pour les injustices du passé et, plus important, de créer des modèles de rôle et de surmonter la discrimination restante et peut-être involontaire.
SYSTRAN	These policies are regarded as a means of offering repairs for the injustices of the past and, more important, of creating models of role and of overcoming remaining and perhaps involuntary discrimination.
Moses	these policies are regarded as a way to offer of repairs for past injustices and , more important , to create a role models and remaining discrimination and perhaps involuntary .
Relearnt	these policies are regarded as a means to offer repairs for the past injustices and , more important , creating role models and overcome remaining discrimination and perhaps involuntary .
REF	These policies are seen as a way of offering reparation for past injustices and, more importantly, for creating role models and for overcoming residual and perhaps involuntary discrimination.

Table 7 : Example outputs for the news domain models (example taken from the **nc-test2007** corpus)

4 Conclusion

The *relearnt* experiment primary goal was to set-up a comparison between three different systems, with equivalent resources. This experiment showed that a statistical translation system may be granted a high BLEU score, even if its translation model was not extracted from corpus. It remains to be seen how this correlates with human judgment (Callison-Burch, 2006), but the detailed error analysis we performed already shows improvements for important categories of errors.

This experiment provided us with some new insight on the strengths and weaknesses of rule-based and phrase-based systems. As an intermediate between a purely corpus-based statistical system and a rule-based system, this setup could benefit from some of the strengths of a phrase-based statistical system, though at the expense of its known drawbacks.

As future work, we may pursue in this direction by exploring the effect of the size of the monolingual corpus used for training the translation model. We may also refine the model by using the target side of the parallel training data when building the language model corpus (to avoid a mismatch of vocabularies) and also combine such a model with the translation model(s) trained on whatever parallel data is available. This would then be interesting to compare this strategy with the corpus-based-only strategies that make use of smaller in-domain parallel corpora.

References

Chris Callison-Burch, Miles Osborne and Philipp Koehn, 2006. *Re-evaluating the Role of Bleu in Machine Translation Research.* In Proceedings of EACL-2006

L. Dugast, J. Senellart and P. Koehn. *Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. Proc. 2nd ACL Workshop on Statistical Machine Translation*, pp. 220-223, June 2007.

Philipp Koehn &al. *Moses: Open Source Toolkit for Statistical Machine Translation*, ACL 2007, demonstration session

Philipp Koehn and Josh Schroeder. Experiments in Domain Adaptation for Statistical Machine Translation, ACL Workshop on Statistical Machine Translation 2007

Holger Schwenk, Jean-Baptiste Fouet and Jean Senellart. *First Steps towards a general purpose French/English Statistical Machine Translation System.* Submitted at the 3rd ACL Workshop on Statistical Machine Translation, 2008

Jean Senellart. 2006. *Boosting linguistic rule-based MT system with corpus-based approaches.* In Presentation. GALE PI Meeting, Boston, MA

M. Simard, C. Goutte, and P. Isabelle. *Statistical Phrase-based Post-Editing. Proc. HLT-NAACL*, pp. 508-515, April 2007.

Simard Michel & al. 2007. *Rule-based Translation With Statistical Phrase-based Post-editing*. In Proceedings of WMT07

Nicola Ueffing, Jens Stephan, Evgeny Matusov, Loïc Dugast, George Foster, Roland Kuhn, Jean Senellart, and Jin Yang *Tighter Integration of Rule-based and Statistical MT in Serial System Combination.* Submitted