Statistical Post-Editing on SYSTRAN's Rule-Based Translation System

Loïc Dugast, Jean Senellart
SYSTRAN SA
La Grande Arche
1, Parvis de la Défense
92044 Paris La Défense Cedex
France
dugast@systran.fr
senellart@systran.fr

Philipp Koehn
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
United Kingdom
pkoehn@inf.ed.ac.uk

Abstract

This article describes the combination of a SYSTRAN system with a "statistical post-editing" (SPE) system. We document qualitative analysis on two experiments performed in the shared task of the ACL 2007 Workshop on Statistical Machine Translation. Comparative results and more integrated "hybrid" techniques are discussed.

1 Introduction

The evolution of SYSTRAN's architecture over the last years has been to « open » the system to enable interaction between the internal system's rules and the external input – see Senellart (2003), Attnas et al. (2005). Based on this architecture, several directions are explored to introduce the use of « corpus-based » approaches at several levels of the process:
- use of corpus-based tools to validate and enrich linguistic resources (detection of forbidden sequences, bilingual terminology extraction), - automatic recognition of the text domain, - use of a corpus-based decision mechanism within « word boundary » (Chinese word identification), disambiguation… - use of word sense disambiguation techniques – and the use of a language model in the generation phase to select alternative translations, prepositions, and local reordering (adjective positioning).

These tools have been presented in Senellart (2006) and most of them will be integrated in SYSTRAN version 7 systems.

Independently, two experiments were carried out for the shared task of the ACL 2007 Workshop on Statistical Machine Translation to combine a raw SYSTRAN system with a statistical post-editing (SPE) system. One experiment was run by NRC using the language pair English<>French in the context of « Automatic Post-Edition » systems using the PORTAGE system as described in Simard et al. (2007). The second experiment based on the same principle was run on the German>English and Spanish>English[1] language pairs using the Moses system (Koehn et al. 2007). The objective was to train a SMT system on a parallel corpus composed of SYSTRAN translations with the referenced source aligned with its referenced translation.

Beyond both (a) the huge (and expected) improvement of the BLEU score for the combined system compared to raw translation output (for German-English, around 10 BLEU points for the Europarl test set of WMT2007) and (b) the (expected) corresponding improvement of the translation fluency, we provide qualitative analysis on the contributions (positive and negative) of the SPE layer imposed on the SYSTRAN translation output in this paper. For this analysis we classifiy the different types of "post-editing" changes and point out the alternative isolated statistical components that could achieve the same results.

We conclude with two possible approaches: breaking down the "statistical layer" into different components/tools each specialized in a narrow and accurate area, or refining this global SPE approach in order to introduce linguistic constraints.

[1] The Moses model was trained following the recommendations for the baseline system of WMT 2007.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 220–223,
Prague, June 2007.

2 The SYSTRAN System

Covering 80 language pairs for 22 different source languages, SYSTRAN powers almost all major portals (Google, Yahoo!, BabelFish, Apple, Worldlingo, …) with machine translation services through URL translations or translation "boxes" (estimated traffic: over 40 million sentence translations and over 10 million web page translations per day).

Customized systems are used by corporate customers either within a post-editing workflow, or without post-editing for the translation of technical Knowledge Bases.

SYSTRAN engines are also available as desktop applications through "plugins" or within post-editing tools. The same engines are also available on ultra-light architectures such as for PDA devices.

The SYSTRAN system is traditionally classified as a "rule-based" system and its design – which has been in constant evolution - has, over the years, always been driven by pragmatic considerations – progressively integrating most of the available productive techniques. As such, it is difficult to classify SYSTRAN and simply describe its architecture. However, the evolution of the SYSTRAN system is governed by the following principles:

- provide a deterministic output : it is possible to easily explain the translation results for a specific sentence and change the rule
- incremental translation quality: the more important evaluation criterion for mature systems is to perform a comparative evaluation of translation output between two consecutive versions. Since it is impossible to guarantee 0 regressions in linguistic development, 8 improvements for 1 degradation defines the acceptance criterion for a linguistic patch.

Crucial components of the SYSTRAN system are the linguistic resources for each language/language pair ranging from 100k to 800k entries. Such "entries" should be understood as both simple or multiword "lexical entries" but also as customized disambiguation rules.

In this context (continuous integration of new techniques in SYSTRAN engines, adhering to determinism and incrementability), over the last three years one major evolution within SYSTRAN has been to make use of available corpora - statically through extraction/learning/validation tools such as:

- Dictionary improvement using a monolingual corpus: new terms/entities/terminology extraction (n-grams based on linguistic patterns);

and dynamically through corpus-based decision algorithms such as:

- Word sense disambiguation
- Use of a language model to select alternative translations, determiner choice, and local controlled reordering – like multiple adjective sequences.

In the following section, we present a qualitative review of the SYSTRAN+SPE output and analyze how the different contributions relate to each specific effort.

3 Experimental Results & Linguistic Evaluation

Based on the data from these two experiments: SYSTRAN+PORTAGE (En<>Fr), and SYSTRAN+Moses (De>En, Es>En), we performed linguistic evaluations on the differences between raw SYSTRAN output and SYSTRAN+SPE output. The evaluation for En<>Fr was performed on the News Commentary test 2006 corpus, while the evaluations for De>En, and Es>En were performed on the Europarl test 2007 corpus.

3.1 Impact

The first observation is the impact of the SPE on the SYSTRAN output. Table 1 displays the WCR (Word Change Rate[2]) and the ratio of sentences impacted by the statistical post-editing. It is interesting to note that the impact is quite high since almost all sentences were post-edited. On the other hand, the WCR of SYSTRAN+SPE is relatively small – as this clearly relates to post-editing and not a complete reshuffling of the translation. The same insight is reinforced when reviewing a comparator (see Table 2) – the SYSTRAN+SPE output

[2] Word Change Rate is computed similarly to the Word Error Rate, with regard to the SYSTRAN output.

is "reasonably" close to the raw SYSTRAN output, and the SPE output structure is completely based on the SYSTRAN output.

	Word Change Rate	Impact (ratio of sentences impacted)
SYSTRAN+PORTAGE En>Fr (nc devtest 2006)	0.33	98%
SYSTRAN+PORTAGE Fr>En (nc devtest 2006)	0.23	95%
SYSTRAN+Moses De>En (nc test 2007)	0.35	100%
SYSTRAN+Moses Es>En (nc test 2007)	0.31	99%

Table 1 - Impact of SPE on raw translation output

Source :En>Fr,De>En,Es>en	SYSTRAN	SYSTRAN +SPE
Monetary policy can be used to stimulate an economy just as much as fiscal policy, if not more, in election years, which politicians will always want to do.	La politique monétaire peut être employée pour stimuler une économie juste comme beaucoup que la politique fiscale, sinon plus, en années d'élection, que les politiciens voudront toujours faire.	La politique monétaire peut être utilisée pour stimuler l'économie, tout comme la politique fiscale, pour ne pas dire plus, dans les années d'élection, que les hommes politiques voudront toujours faire.
Fortschritte der 12 Bewerberländer auf dem Weg zum Beitritt	Progress of the 12 applicant countries on the way to the entry	Progress of the 12 candidate countries along the road to accession
En una perspectiva a más largo plazo, habrá una moneda única en todo el continente.	In a perspective to more long term, there will be a unique currency in all the continent.	In a more long-term perspective, there will be a single currency for the whole continent.

Table 2 - Comparison of source, SYSTRAN, and SYSTRAN+SPE: the output is "reasonably close" – and clearly preserves SYSTRAN's translation structure

3.2 Linguistic Categorization of Different Post-Editing Changes

To classify the types of "post-editing" changes brought by the SPE system, we define the following criteria:

- `termchg` – changes related to lexical changes.
  - `termchg_nfw` – word not translated by SYSTRAN generating a translation with SPE.
  - `termchg_term` – slight terminology change preserves part of speech and meaning. Most of the time changes improve fluency by selecting the appropriate terminology. (e.g. *politicians→politiciens* vs. the more commonly used "*hommes politiques*").
  - `termchg_loc` – multiword expression/locution change (*the same is true→Le même est vrai* vs. *C'est également vrai*)
  - `termchg_mean` – lexical modification altering the meaning of the sentences, by changing the part of speech of the word, or by selecting a completely different meaning for a given word. (*Despite occasional grumbles→En dépit des grognements occasionnels* vs. *En dépit des maux économiser*)
- `gram` – changes related to grammar
  - `gram_det` – change in determiner (*on political commitments→sur des engagements politiques* vs. *sur les engagements politiques*)
  - `gram_prep` – change in preposition (*across the Atlantic→à travers l'atlantique* vs. *de l'autre côté de l'atlantique*)
  - `gram_pron` – change in pronoun
  - `gram_tense` – change in tense (*should not be hidden→ne devraient…* vs. *ne doivent…)*
  - `gram_number/gram_gender` – change in number/gender – often reflecting lack of agreement
  - `gram_other` – other grammatical changes
- `punct/digit/case` – change in punctuation, case, or numbers
- `wordorder_local` – change in local word order
- `wordorder_long` – change in word order (long distance)
- `style` – change in "style" (*justifying→justifiant* vs. *ce qui justifie*)

A detailed count of the number of improvements (*#improv*), degradations (*#degrad*) and equivalents (*#equiv*) related to each category performed for a sample corpus (100 sentences each) for En>Fr, De>En and Es>En systems, and related results are reported in the following tables[3]:

	SYSTRAN PORTAGE En>Fr	SYSTRAN Moses De>En	SYSTRAN Moses Es>En
`termchg all`	+22%	+46%	+46%
`termchg_nfw`	0%	+3%	+1%
`termchg_term`	**+19%**	**+42%**	**+45%**
`termchg_loc`	**+8%**		
`termchg_mean`	**-6%**		
`gram all`	+2%	+4%	+12%
`gram_det`	**14%**	+2%	+4%
`gram_prep`	2%	+1%	+5%
`gram_pron`	-1%	+1%	+4%
`gram_tense`	**-4%**	+1%	-0%
`gram_number`	0%	None	None
`gram_gender`	**-4%**	n/a	n/a
`gram_other`	**-1%**	None	None
`punct/digit/case`	1%	**-1%**	**-1%**
`wordorder_short`	-1%	+1%	+1%
`wordorder_long`	0%	None	+1%
`style`	1%	+3%	+2%

Table 3 - Relative improvements brought by the SPE system: (*#improv-#degrad*)/∑*#modif*

	#improv	*#degrad*	*#improv / #degrad*	*#equiv*
`termchg all`	90	32	3	33
`termchg_nfw`	1	0		0
`termchg_term`	**59**	7	**8**	29
`termchg_loc`	15	1	**15**	1
`termchg_mean`	15	**24**	1	3
`gram all`	44	38	1	8
`gram_det`	**20**	3	**7**	4
`gram_prep`	**12**	9	1	1
`gram_pron`	0	1	0	2
`gram_tense`	2	**8**	0	0
`gram_number`	4	**4**	1	0
`gram_gender`	2	**8**	0	0

[3] Manual evaluations for De>En and Es>En should not be compared with the results for En>Fr, as both corpus and evaluation criteria differ.

gram_other	4	5	1	1
punct/digit/case	8	7	1	1
wordorder_short	0	1	0	0
wordorder_long	0	0		0
style	3	1	3	1

Table 4 - Details on *#improv*, *#degrad*, *#equiv* for each category for SYSTRAN PORTAGE En>Fr

3.3 Analysis of Results

The figures from the previous section provide very useful information that requires deeper analysis, the most obvious of which follow:

- As is, this basic integration does not meet the acceptance criterion "8 improv. for 1 degrad."
- The most improved category is the "termchg" which corresponds to a local choice of word sense or alternative translation of words and locutions. In this category, the main source degradation stems from the "termchg_mean" category. This category covers changes of lexical unit parts of speech.
- In grammatical categories, productive categories are "gram_det" and "gram_prep" but the improvement/degradation ratio for this last category is very low (it shows global improvements but there are many unacceptable degradations).
- As expected, no "long-distance" restructuring is observed and local reordering is negative for En>Fr and relatively negligible for other language pairs.
- For the French target, morphology is a major issue (accounts for 25% of degradations). This was also expected since no mechanism in the SPE provides any control over the morphology.

4 Conclusions

The SYSTRAN+SPE experiments demonstrate very good results – both on automatic scoring and on linguistic analysis. Detailed comparative analysis provides directions on how to further improve these results by adding "linguistic control" mechanisms. For SPE, we would, for instance, add linguistic constraints in the decoding process, knowing that the structure/linguistic information could be made available in the translation output.

Beyond the scope of these experiments, our results set a baseline to compare with other more sophisticated/integrated "rules and statistics" combination models.

In particular, the most improved categories observed in these experiments confirm that our current development direction for integrating data-driven mechanisms within translation engines (especially for word sense disambiguation, for the selection of alternative translations or for specific local phenomena like determination) should converge on the same results while preventing associated degradations. Also, the high score reached by the "termchg_loc" category substantiates the need to continue exploiting phrase tables built on parallel corpora to learn new terminology.

Acknowledgments

We would like to thank Michel Simard, Roland Kuhn, George Foster and Pierre Isabelle from NRC, Canada for their collaboration on this work (Simard et al. 2007).

References

Attnäs (M.), Senellart (P.) and Senellart (J.). 2005. *Integration of SYSTRAN MT systems in an open workflow*. Machine Translation Summit, Phuket, Thailand.

Philipp Koehn & al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. To appear at ACL2007, Prague.

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

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

Jean Senellart, & al. 2003. *XML Machine Translation*. In Proceedings of MT-Summit IX.

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