DEPARTAMENTO DE SISTEMAS
INFORMÁTICOS Y COMPUTACIÓN

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Interactive Post-Editing in Machine Translation

MASTER'S THESIS

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Author: Miguel Domingo Ballester

Supervisor: Francisco Casacuberta Nolla

September 25, 2015

Outline

Table of Contents

Goal

The goal of this thesis is to develop a methodology that increases the translation quality of a given machine translation system, and reduces the effort a human agent would need to make to correct the translations of that system (post-editing), by interacting with the human agent on a statistical post-editing system.

Table of Contents

Machine Translation

Machine Translation (MT) is a discipline that researches the use of computer software to automatically translate from one language to another.

Computer Assisted Translation (CAT) seeks to provide human translators with as many tools as possible in order to facilitate their work through computer software.

There are many approaches to machine translation, but in this work we have used:

- Rule-Based Machine Translation (RBMT).
- Statistical Machine Translation (SMT).

Statistical Machine Translation

Given a sentence $\mathbf{x}$ in a source language, the problem in MT is to find its corresponding translation $\mathbf{y}$ in a target language (Brown et al., 1993):

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y}|\mathbf{x})$$

In practice, this is often combined into a *log-linear model* for $Pr(\mathbf{y}|\mathbf{x})$ (Och & Ney, 2002):

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \left\{ \sum_{n=1}^{N} \lambda_n \cdot log(f_n(\mathbf{y}, \mathbf{x})) \right\}$$

where $f_n(\mathbf{y}, \mathbf{x})$ can be any model that represents an important feature for the translation; N is the number of models (target language model, direct phrase-based model, inverse model...); and λ_n are the weights of the log-linear combination.

Domain Adaptation

Often, there is a mismatch between the target domain of an SMT system (*in-domain*), and the domain from which training data are available (*out-of-domain*).

Domain Adaptation aims to improve performance of systems trained with out-of-domain data.

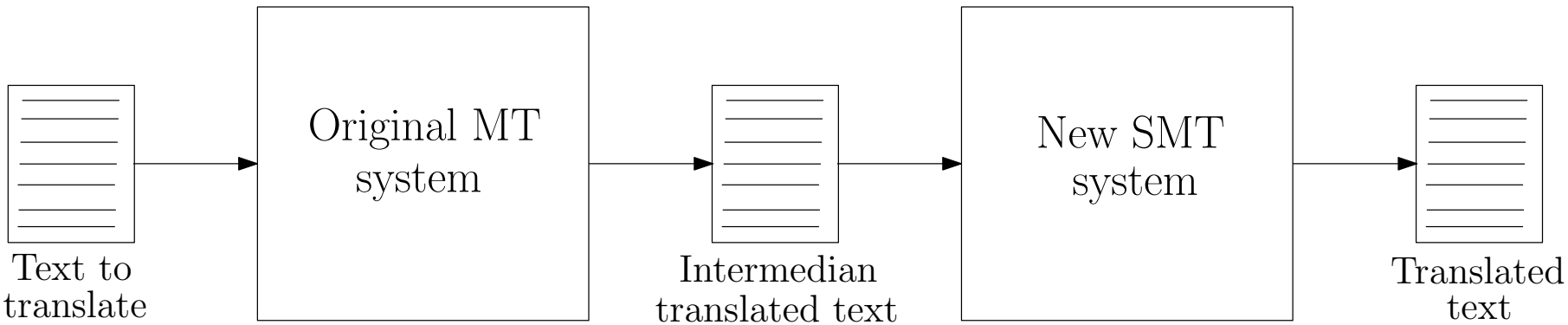
Post-Editing

The current state of the art in MT is far from being good enough, and a manual correction is needed to improve the quality of translations.

This manual correction, known as post-editing, is the process of improving a machine-generated translation with a minimum of manual labour (TAUS, 2010).

Statistical Post-Editing

Statistical Post-Editing (SPE) (Simard et al., 2007) is a technique to automatically improve the translation quality of a MT system, by training a new SMT system that "corrects" the translations of that system.



Main research areas:

- Post-editing a **RBMT** system.
- Using SPE as a **domain adaptation** technique.

Interactive Translation Prediction

Interactive Translation Prediction (ITP) (Barrachina et al., 2009) is a methodology to reduce human post-editing effort by interacting with the human agent to obtain the final translation.

source (x): The cough may last for 1-2 months or longer
desired translation ($\hat{\mathbf{y}}$): La tos puede durar 1-2 meses o más tiempo

IT-0	**p** $\mathbf{s_h}$	 La tos puede durar tres meses o más
IT-1	**p** *w* $\mathbf{s_h}$	La tos puede durar 1-2 meses o más
IT-2	**p** *w* $\mathbf{s_h}$	La tos puede durar 1-2 meses o más tiempo
END	**p**	La tos puede durar 1-2 meses o más tiempo

Interactive Translation Prediction

The crucial step in ITP is the production of the new suffix, whose probability is given by:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s}_\mathbf{h}} Pr(\mathbf{s}_\mathbf{h}|\mathbf{x}, \mathbf{p})$$

where $\mathbf{s}_\mathbf{h}$ is the new suffix, $\mathbf{p}$ is the known prefix, and $\mathbf{x}$ is the source sentence.

Applying Bayes' theorem, this equation can be seen as:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s}_\mathbf{h}} Pr(\mathbf{p}, \mathbf{s}_\mathbf{h}|\mathbf{x})$$

which is similar to the main equation of MT, but with the search procedure limited to those target sentence $\mathbf{y}$ whose prefix is equal to $\mathbf{p}$.

Table of Contents

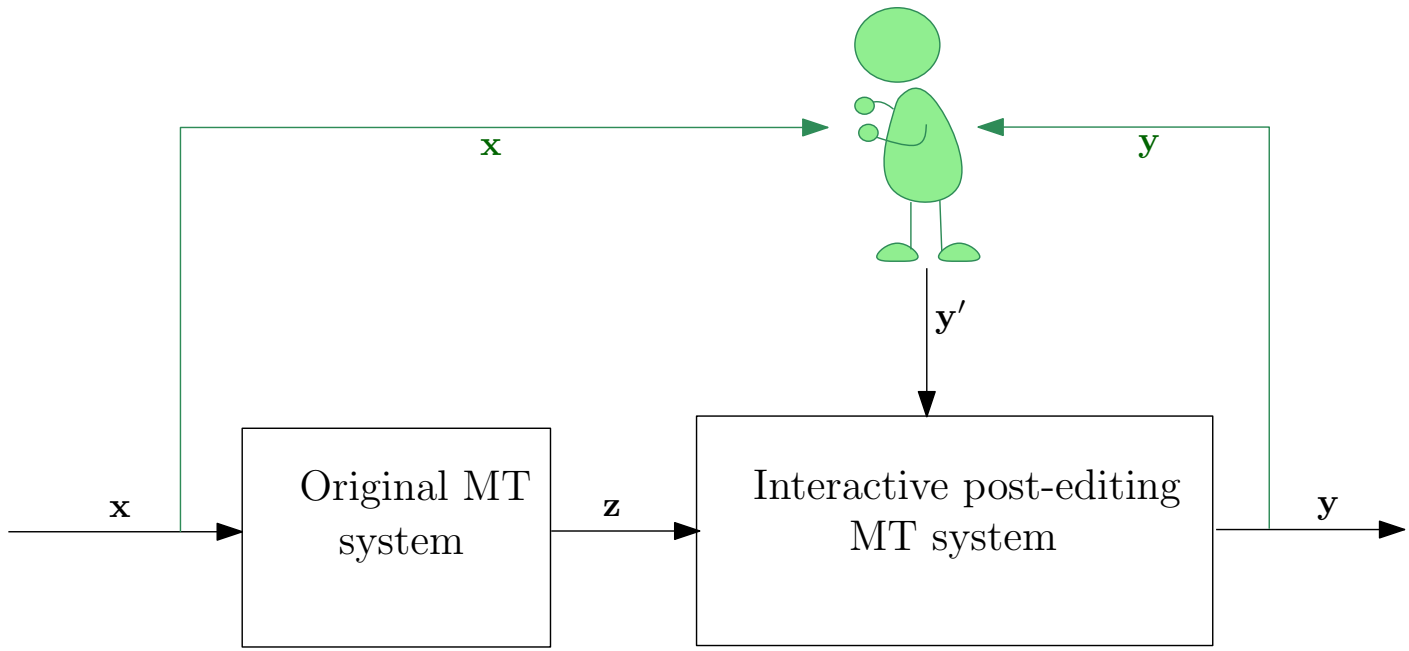
Interactive Post-Editing

In this work we have proposed a methodology to improve translation quality and reduce human post-editing effort. This methodology is the combination of the SPE technique and the ITP methodology.

The main difference of this methodology with SPE, is that we are using an interactive process. And, at the same time, this methodology is different of ITP in that we are applying the interactivity to an statistical post-editing system.

Interactive Post-Editing

Each time a source sentence **x** is fed into the system, the propose methodology performs the following process:



The source sentence is first translated with the original MT system, generating the new source sentence **z**. Then, the new source sentence is fed into the interactive post-editing system which, with the interaction of the user, generates the final target sentence **y**.

Interactive Post-Editing

Like in ITP, the crucial step in this process is the production of the new suffix, whose probability is given by:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s}_\mathbf{h}} Pr(\mathbf{s}_\mathbf{h}|\mathbf{x}, \mathbf{z}, \mathbf{p})$$

where $\mathbf{s}_\mathbf{h}$ is the new suffix, $\mathbf{p}$ is the known prefix, $\mathbf{x}$ is the original source sentence, and $\mathbf{z}$ is the adapted source sentence.

Applying Bayes' theorem, this can be seen as:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s}_\mathbf{h}} Pr(\mathbf{p}, \mathbf{s}_\mathbf{h}|\mathbf{x}, \mathbf{z})$$

For simplicity, in this work we have assumed independence with respect to $\mathbf{x}$:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s}_\mathbf{h}} Pr(\mathbf{p}, \mathbf{s}_\mathbf{h}|\mathbf{z})$$

Table of Contents

Corpora
EMEA

Formed by documents of the *European Medicines Agency* (Tiedemann, 2009).

		English	Spanish
Training	Sentences	292K	
	Running words	5M	5.6M
	Vocabulary	67K	81K
Development	Sentences	9.8K	
	Running words	171K	192K
	Perplexity (3-grams)	30.43	30.76
Test	Sentences	5K	
	Running words	90K	101K
	Perplexity (3-grams)	31.72	31.61

Corpora
EU

Extracted from the *Bulletin of the European Union* (Khadivi & Goutte, 2003).

		English	Spanish
Training	Sentences	212.4K	
	Running words	5.2M	5.8M
	Vocabulary	40.3K	53.8K
Development	Sentences	2K	
	Running words	49.8K	55.8K
	Perplexity (3-grams)	37.42	31.98
Test	Sentences	800	
	Running words	20K	22.8K
	Perplexity (3-grams)	43.82	36.46

Corpora
Europarl

Extracted from the *Proceedings of the European Parliament* (Koehn, 2005).

		English	Spanish
Training	Sentences	1.9M	
	Running words	53.1M	55.5M
	Vocabulary	130K	190K
Development	Sentences	10K	
	Running words	268K	279K
	Perplexity (3-grams)	58.07	60.01

Metrics

- BLEU: Geometric average of the modified n-gram precision multiplied by a factor that penalizes short sentences.
- WER: Minimum number of operations to convert the word strings of the translated sentences, into the word strings of the reference sentences.
- WSR: Number of word strokes a user would need to correct each sentence, normalized by the number of words.
- E-R: Relative difference between WER and WSR.

Experimental Set-Up

- **Post-editing** of a RBMT system (created with general rules).
  - EMEA corpus.
  - EU corpus.
- **Domain adaptation** of a SMT system (trained with europarl corpus).
  - EMEA corpus.
  - EU corpus.

RBMT Experiment
EMEA

This experiment assesses the performance of an SMT system trained with EMEA corpus, a general RBMT system, and the application of the proposed methodology to the RBMT system; when translating the EMEA test.

System	en-es			
	Automatic		Interactive	
	BLEU (%)	WER (%)	WSR (%)	E-R (%)
SMT (EMEA)	58.2	42.3	36.4	14.1
RBMT	17.6	71.6	–	–
+ SPE	**59.5**	**40.3**	**36.2**	**10.2**

RBMT Experiment
EU

This experiment assesses the performance of an SMT system trained with EU corpus, a general RBMT system, and the application of the proposed methodology to the RBMT system; when translating the EU test.

System	en-es			
	Automatic		Interactive	
	BLEU (%)	WER (%)	WSR (%)	E-R (%)
SMT (EU)	**48.6**	**44.6**	**44.0**	**1.3**
RBMT	14.8	70.5	–	–
+ SPE	48.0	44.7	44.5	0.3

Domain Adaptation Experiment
EMEA

This experiment assesses the performance of an SMT system trained with EMEA corpus, an out-of-domain SMT system trained with europarl corpus, and the application of the proposed methodology to the out-of-domain system; when translating the EMEA test.

System	**en-es**			
	Automatic		**Interactive**	
	BLEU (%)	**WER (%)**	**WSR (%)**	**E-R (%)**
SMT (EMEA)	**58.2**	**42.3**	**36.4**	**14.1**
SMT (europarl)	23.9	67.0	71.9	-7.3
+ SPE (EMEA)	56.6	42.6	39.0	8.5

Domain Adaptation Experiment
EU

This experiment assesses the performance of an SMT system trained with EU corpus, an out-of-domain SMT system trained with europarl corpus, and the application of the proposed methodology to the out-of-domain system; when translating the EU test.

System	**en-es**			
	Automatic		**Interactive**	
	BLEU (%)	**WER (%)**	**WSR (%)**	**E-R (%)**
SMT (EU)	**48.6**	**44.6**	**44.0**	**1.3**
SMT (europarl)	34.2	54.6	54.7	-0.2
+ SPE (EU)	46.9	46.2	46.1	0.2

Table of Contents

Conclusions

- We have proposed a methodology to increase translation quality and reduce human post-editing effort.
- We have tested the methodology with different systems and corpora.
- The methodology always improved the original MT system.
- These results weren't better than training a new system with the in-domain corpora.
- We think we are in the right direction to develop a methodology to increase translation quality and reduce human post-editing effort, but we still have some more work to do.

Future Work

- Experiment with more MT systems and more diverse corpora.
- Try an adapted version of a RBMT system instead of just a general version.
- Incorporate a new module into our methodology that enables the possibility to have more that one MT system.
- Try using only a part of the in-domain corpus (from a small part up to the whole corpus) to study the cases in which there is a small quantity of in-domain data.
- Compare our methodology to other domain adaptation techniques.

Table of Contents

Bibliography

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., ... Vidal, E. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, *35*(1), 3–28.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Khadivi, S., & Goutte, C. (2003). *Tools for corpus alignment and evaluation of the alignments (deliverable d4. 9)* (Tech. Rep.). Technical report, TransType2 (IST-2001-32091).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit* (pp. 79–86). Retrieved from `http://mt-archive.info/MTS-2005-Koehn.pdf`

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 295–302).

Simard, M., Goutte, C., & Isabelle, P. (2007). Statistical phrase-based post-editing. In *Proceedings of naacl.*

TAUS. (2010). *Postediting in practice.* (TAUS report `https://www.taus.net/reports/postediting-in-practice`)

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing* (Vol. V, pp. 237–248). Borovets, Bulgaria.