Poor man's lemmatisation for automatic error classification

Maja Popović[1] **Mihael Arčan**[2] **Eleftherios Avramidis**[1]
Aljoscha Burchardt[1] **Arle Lommel**[1]

[1] DFKI – Language Technology Lab, Berlin, Germany
firstname.lastname@dfki.de
[2] Insight Centre for Data Analytics, National University of Ireland, Galway
mihael.arcan@insight-centre.org

Abstract

This paper demonstrates the possibility to make an existing automatic error classifier for machine translations independent from the requirement of lemmatisation. This makes it usable also for smaller and under-resourced languages and in situations where there is no lemmatiser at hand. It is shown that cutting all words into the first four letters is the best method even for highly inflective languages, preserving both the detected distribution of error types within a translation output as well as over various translation outputs.

The main cost of not using a lemmatiser is the lower accuracy of detecting the inflectional error class due to its confusion with mistranslations. For shorter words, actual inflectional errors will be tagged as mistranslations, for longer words the other way round. Keeping all that in mind, it is possible to use the error classifier without target language lemmatisation and to extrapolate inflectional and lexical error rates according to the average word length in the analysed text.

1 Introduction

Future improvement of machine translation (MT) systems requires reliable automatic evaluation and error classification tools in order to minimise efforts of time and money consuming human classification. Therefore automatic error classification tools have been developed in recent years (Zeman et al., 2011; Popović, 2011) and are being used to facilitate the error analysis. Although these tools are completely language independent, for obtaining a precise error distribution over classes a lemmatiser for the target language is required. For the languages strongly supported in language resources and tools this does not pose a problem. However, for a number of languages a lemmatiser might not be at hand, or it does not exist at all. This paper investigates possibilities for obtaining reasonable error classification results without lemmatisation. To the best of our knowledge, this issue has not been investigated so far.

© 2015 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

2 Motivation and explored methods

We investigate the edit-distance i.e. word error rate (WER) approach implemented in the Hjerson tool (Popović, 2011), which enables detection of five error categories: *inflectional errors*, *word order errors*, *missing words* (omissions), *extra words* (additions) and *lexical errors* (mistranslations). For a given MT output and reference translation, the classication results are provided in the form of the five error rates, whereby the number of errors for each category is normalised over the total number of words.

The detailed description of the approach can be found in (Popović and Ney, 2011). The starting point is to identify actual words contributing to the Word Error Rate (WER), recall (reference) error rate (RPER) and precision (hypothesis) error rate (HPER). The WER errors are marked as substitutions, deletions and insertions. Then, the lemmas are used: first, to identify the inflectional errors – if the lemma of an erroneous word is correct and the full form is not. Second, the lemmas are also used for detecting omissions, additions and mistranslations. It is also possible to calculate WER

Method	
full	The visit will reach its peak in the afternoon .
$lemma$	The visit will reach its peak in the afternoon .
$4let$	The visi will reac its peak in the afte .
$2thirds$	Th vis wi rea it pe in th aftern .
$stem$	The visi wil rea its pea in the afternoo .
full	President is receiving the Minister of Finance .
$lemma$	President be receive the Minister of Finance .
$4let$	Pres be rece the Mini of Fina .
$2thirds$	Presid is receiv th Minis of Fina .
$stem$	Presiden is receiv the Minist of Financ .

Table 1: Examples for each of the word reduction methods.

based on lemmas instead of full words in order to increase the precision with regard to human error annotation, which makes the algorithm even more susceptible to possible lack of lemmas.

If the full word forms were used as a replacement for lemmas, it would not be possible to detect any inflectional error thus setting the inflectional error rate to zero, and noise would be introduced in omission, addition and mistranslation error rates. Therefore, a simple use of the full forms instead of lemmas is not advisable, especially for the highly inflective languages. The goal of this work is to examine possible methods for processing of the full words in a more or less simple way in order to yield a reasonable error classification results by using them as a replacement for lemmas. Following methods for word reduction are explored:

- first four letters of the word ($4let$)

  The simplest way for word reduction is to use only its first n letters. The choice of first four letters has been shown to be successful for improvement of word alignments (Fraser and Marcu, 2005), therefore we decided to set n to four.

- first two thirds of the word length ($2thirds$)

  In order to take the word length into account, the words are reduced to 2/3 of their original length (rounded down).

- word stem ($stem$)

  A more refined method which splits words into stems and suffixes based on harmonic mean of their frequencies is used, similar to the compound splitting method described in (Koehn and Knight, 2003). The suffix of each word is removed and only the stem is preserved. For calculation of stem and suffix frequencies, both the translation output and its corresponding reference translation are used.

Examples of two English sentences processed by each of the methods is shown in Table 1.

The methods are tested on various distinct target languages and domains, some of the languages being very morphologically rich. Detailed description of the texts can be found in the next section.

3 Experiments and results

The two main objectives of automatic error classifier are:

- to estimate the error distribution within a translation output
- to compare different translation outputs in terms of error categories

Therefore we tested the described methods for both these aspects by comparing the results with those obtained when using lemmatised words, i.e. we used the error rates obtained with lemmas as the "reference" error rates. The best way for the assessment would be, of course, a comparison with human error classification. Nevertheless, this has not been done for two reasons: first, the original method using lemmas is already thoroughly tested in previous work (Popović and Ney, 2011) and is shown to correlate well with human judgements. Second, human evaluation is resource and time-consuming.

The explored target languages in this work are English, Spanish, German, Slovenian and Czech

originating from news, technical texts, client data of Language Service Providers, pharmaceutical domain, Europarl (Koehn, 2005), as well as the OpenSubtitles[1] spoken language corpus. In addition, one Basque translation output from technical domain has been available as well. The publicly available texts are described in (Callison-Burch et al., 2011), (Specia, 2011) and (Tiedemann, 2012). The majority of translation outputs has been created by statistical systems but a number of translations has been produced by rule-based systems. It should be noted that not all target languages were available for all domains, however the total amount of texts and the diversity of languages and domains are sufficient to obtain reliable results – about 36000 sentences with average number of words ranging from 8 (subtitles) through 15 (domain-specific corpora) up to 25 (Europarl and news) have been analysed.

Lemmas for English, Spanish and German texts are generated using TreeTagger,[2] Slovenian lemmas are produced by the Obeliks tagger (Grčar et al., 2012), and Czech texts are lemmatised using the COMPOST tagger (Spoustová et al., 2009).

It should be noted that all the reported results are calculated using WER of lemmas (or corresponding substitutions) since no changes related to lemma substitution techniques were observed in comparison with the use of the standard full word WER.

3.1 Error distributions within a translation output

Our first experiment consisted of calculating distributions of five error rates within one translation output using all word reduction methods described in Section 2 and comparing the obtained results with the reference distributions of error rates obtained using lemmas. The results for three distinct target languages are presented in Table 2: English as the least inflective, Spanish having very rich verb morphology, and Czech as generally highly inflective.

Reference distributions are presented in the first row, followed by the investigated word reduction methods; in the last row the results obtained using full words are shown as well, and the intuitively suspected effects can be clearly seen: no inflectional errors are detected, and the vast majority of them are tagged as lexical error (mistranslation). Furthermore, it is confirmed that the variations in word order errors, omissions and additions are small, whereas the most affected error classes are inflections and mistranslations.

As for different target languages, in the English output the differences between the error rates are small for all error classes, but for the more inflected Spanish text and the highly inflected Czech text the situation is fairly different: $4let$ distribution is closest to the reference lemma error distribution, whereas $2thirds$ and $stem$ distributions are lying between the lemma and the full word distributions. In addition, it can be observed that the $stem$ method performs better than the $2thirds$ method.

In Table 3, the parts of the reference translations from Table 1 containing inflectional errors are shown together with the corresponding parts of the translation output in order to better understand the different performance of the methods. Each of the sentences contains one (verb) inflectional error. The first error, "receives" instead of "receiving", is correctly detected by all methods. The second one, "reached" instead of "reach" is correctly tagged by all methods except by $2thirds$ because the reduced word forms are not the same in the translation and in the reference. The $stem$ method often exhibits the same problem, however less frequently.

3.2 Comparing translation outputs

For the comparison of different translation outputs, only the $4let$ method has been investigated because it produces the best error distributions (closest to those obtained by lemmas) and it is also the simplest to perform.

Figure 1 illustrates the results for the two highly inflectional languages, namely Slovenian (above) and Czech (below). Slovenian translations originating from six statistical MT systems (dealing with three different domains and two source languages) and Czech outputs produced by four different MT systems have been analysed. Only the two most critical error classes are presented, namely inflectional (left) and lexical (right) error rates – for other error categories no significant performance differences between the reduction methods were observed.

For the Slovenian translations, the correlation between $4let$ and reference lemma system rankings is 1, both for the inflectional and for the lexical error rates. The same applies to Czech lex-

[1] http://www.opensubtitles.org/

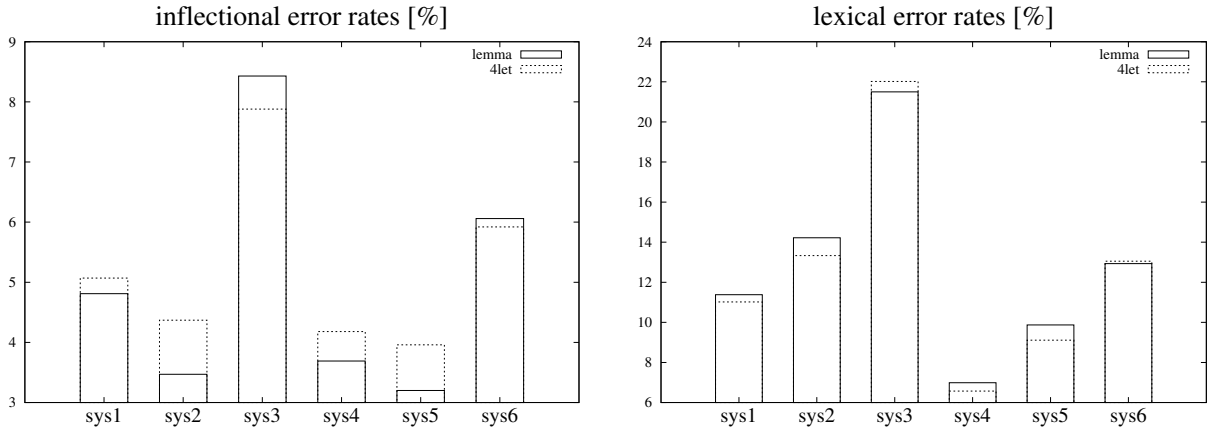
[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Target Language	Method	Error Rates [%] infl	order	miss	add	lex
English	*lemma* (ref)	1.5	7.6	5.2	3.0	8.7
	4let	1.9	7.6	5.2	3.0	8.2
	2thirds	0.9	7.5	5.3	3.0	9.3
	stem	1.2	7.6	5.3	3.0	9.0
	full	0	7.6	5.4	3.1	10.1
Spanish	*lemma* (ref)	4.6	6.4	5.9	3.6	13.5
	4let	4.0	6.6	6.0	3.6	13.9
	2thirds	2.6	6.4	6.0	3.5	15.5
	stem	3.1	6.6	6.1	3.6	14.8
	full	0	6.7	6.1	3.6	17.9
Czech	*lemma* (ref)	10.4	10.6	7.1	7.6	36.4
	4let	10.0	10.8	7.0	7.7	36.9
	2thirds	5.6	11.0	6.8	7.6	41.4
	stem	7.2	10.9	7.0	7.7	39.7
	full	0	11.3	6.8	7.6	47.1

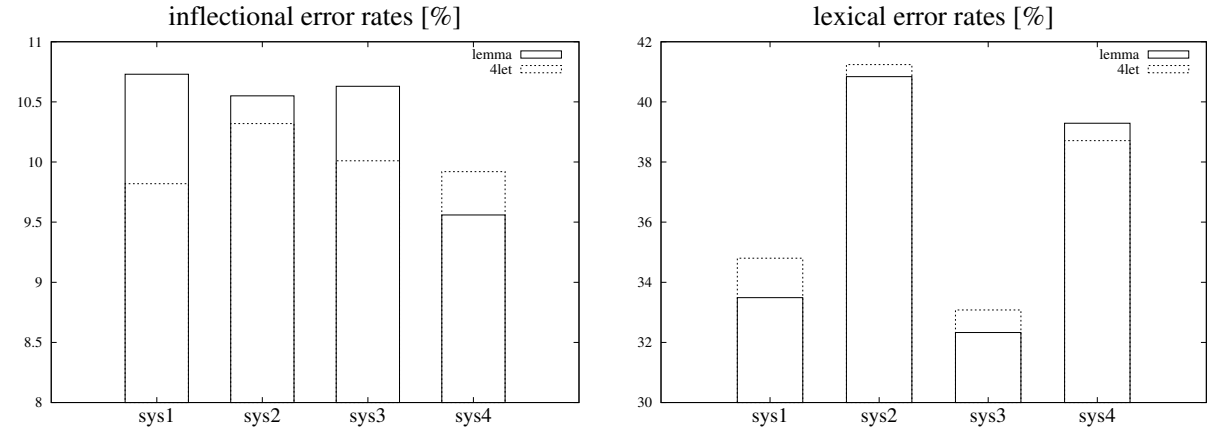
Table 2: Comparison of error rates obtained by each of the described word reduction methods with the reference lemma error rates for three translation outputs: English (above), Spanish (middle) and Czech (below). Error rates using full words as lemma replacement are shown as well, illustrating why this method is not recommended.

Method	Reference translation	MT output
full	The visit will reach	Visit reached
lemma	The visit will *reach*	Visit *reach*
4let	The visi will *reac*	Visit *reac*
2thirds	Th vis wi *rea*	Vis *rea*
stem	The visi will *rea*	Vis *rea*
full	President is receiving	President receives
lemma	President be *receive*	President *receive*
4let	Pres be *rece*	Pres *rece*
2thirds	Presid is **receiv**	President **recei**
stem	Presiden is *receiv*	President *receiv*

Table 3: Illustration of the main problem for inflectional error detection: if the reduced word form is not exactly the same in the reference and in the translation output (bold), the error will not be tagged as inflectional. This phenomenon occurs most frequently for the *2thirds* method, therefore this method exibits the poorest performance.



(a) Comparing six Slovenian MT outputs



(b) Comparing four Czech MT outputs

Figure 1: Comparison of translation outputs for highly inflective languages based on the two most critical error classes, i.e. inflectional (left) and lexical errors (right) – six Slovenian (above) and four Czech (below) translation outputs. Reference lemma error rates are presented by full lines, $4let$ error rates by dashed lines.

ical error rates, but not for the Czech inflections though: lemma method ranks the error rates (from highest to lowest) *1, 3, 2, 4* whereas the $4let$ ranking is *2, 3, 4, 1*. However, the important fact is that the relative differences between the systems are very small for inflectional errors; all the systems contain a high number of inflectional errors (between 9.6 and 10.8%), whereas the absolute differences between the systems range only between 0.2 and 1%. This means that the $4let$ method is generally well capable of system comparison, but it is not able to capture very small relative differences correctly.

3.3 Analysis of confusions

In previous sections it is shown that the $4let$ method, despite certain disadvantages, is well capable to substitute the lemmas both for estimating error distributions within an output as well as for comparing error rates across the translation outputs. However, an important remaining question is: what is exactly happening? Results presented in previous sections indicate that a number of inflectional errors is substituted by lexical errors. However, they also show that the $4let$ inflectional error rates sometimes are lower and sometimes higher than the lemma-based ones, thus indicating that not only a simple substitution of inflectional errors by mistranslations is taking place.

In order to explore these underlying phenomena, accuracies and confusions between error classes are calculated and confusion matrix is presented in Table 4. Since there are practically no variations in reordering error rates, the confusions are presented only for inflections, additions[3] and lexical errors.

As a first step, the confusions are calculated for all merged texts and the results are presented in the first row. It is confirmed that the low accuracy of the inflections and their confusions with mistranslations are indeed the main problems, however there is a number of reverse confusions, i.e. certain mistranslations are tagged as inflectional errors. Apart from that, there is also certain amount of confusions between inflections and additions.

Since some of the used reference translations were independent ("free") human translations and some were post-edited translation outputs, we separated the texts into two sets and calculated confusions for each one. Nevertheless, no important differences could be observed, as it can be seen in the corresponding rows in Table 4.

The next step was to analyse each of the target languages separately, and the results are presented further below in the table. Although the numbers are more diverse, all the important phenomena are practically same for all languages, namely low accuracy of inflections due to confusion with mistranslations. Only for the Basque translation the percentage is similar for confusions in both directions.

Last step was division of texts into written text and spoken language transcriptions, and, contrary to the other set-ups, several notable differences were observed. First of all, the accuracy of inflections is significantly lower for spoken language, and the percentage of confusions with mistranslations is much higher. On the other hand, in written text much more mistranslations are substituted by inflections.

3.3.1 Word length effects

The differences between written and spoken language, together with the observations about Basque where the words can be very long, showed that the word length is an important factor which is neglected by the simple cutting of words into first four letters. The inflections of very short words such as articles and auxiliary verbs cannot be captured, and some long words which are not related at all can be easily tagged as inflectional errors only because they share the first four characters – see Table 5. Furthermore, *reception*, *receipt*, *recent* and *receiver* all share first four letters and could possibly be tagged as inflectional error. On the other hand, such coincidences are not very frequent and therefore there are less substitutions of lexical errors. We calculated the average lengths of words for which each of the two substitution types occur, and obtained an average word length of 3.44 for inflection→mistranslation substitution and 8.64 for the reverse one.

Neglecting the word length by the $4let$ method was the reason to explore the other two methods ($2thirds$ and $stem$) in the first place. However, they produced significantly worse error distributions due to the often inconsistent word cutting. Since the $stem$ method could be potentially improved (contrary to the $2thirds$ method), we analysed its confusions and compared with those of the $4let$ method in order to better understand the differences. The confusions for all merged translation

[3]The situation regarding omissions is analogous to the one regarding additions.

$4let$	infl	infl→lex	infl→add	lex	lex→infl	add	add→infl
Overall	**57.1**	**36.0**	5.6	89.5	8.0	88.9	8.0
Reference	56.2	37.4	5.5	90.5	7.2	90.4	3.7
Post-edit	57.9	34.3	5.8	87.5	9.8	86.8	6.1
English	47.1	46.8	5.5	93.2	5.4	94.7	2.8
Spanish*	55.6	35.2	5.3	89.4	8.6	91.8	2.6
German	43.2	47.0	7.3	87.9	8.5	84.9	8.1
Slovenian	51.6	41.8	6.2	91.9	6.2	86.7	2.1
Czech*	66.3	28.4	5.2	90.0	7.3	81.3	6.3
Basque*	79.2	**16.4**	3.8	84.0	**13.4**	86.3	5.6
Written	65.7	27.4	5.0	**87.0**	**10.1**	87.6	**6.4**
Spoken	**44.4**	**49.2**	6.0	94.1	4.6	89.7	1.6

Table 4: Accuracies and confusions between reference lemma error categories and those obtained by the $4let$ method; for all texts (Overall), separately for post-editions and for references, separately for each target language, and separately for written and spoken language.

Method	Reference translation	MT output
full	There were ergonomic problems .	There was ergonomische problems .
$lemma$	There be ergonomic problem .	There be_{infl} $ergonomische_{lex}$ problem .
$4let$	Ther were ergo prob .	Ther was_{lex} $ergo_{infl}$ prob .

Table 5: Illustration of the word length problem for the $4let$ method: inflectional errors for short words (*were/was*) are impossible to detect and are considered as lexical errors; on the other hand, a lexical error (untranslated German word *ergonomische*) is tagged as inflectional error because it shares first four letters with the reference translation *ergonomic*.

	Method	infl	infl→lex	infl→add	lex	lex→infl	add	add→infl
Overall	$4let$	**57.1**	**36.0**	**5.6**	89.5	8.0	88.9	8.0
	$stem$	48.4	44.2	6.0	**94.2**	**4.8**	**89.7**	**4.3**

Table 6: Comparison of overall $4let$ and $stem$ accuracies and confusions.

outputs (Overall) presented in Table 6 show that the $stem$ method is better in avoiding substitutions of mistranslations and additions with inflections, but the problem with low inflection error accuracy is worse. One possible reason is that the stem and the suffix frequencies are estimated from the very small amount of data (only the reference and the translation output) and therefore is often not able to perform consistent cuttings for all words. This method should be investigated in future work, trained on the large target language corpus as well as in combination with the $4let$ method.

4 Conclusions and Future Work

The experiments presented in this paper show that it is possible to use an existing automatic error classifier without target language lemmas. It is shown that cutting all words into first four letters is the best method even for highly inflective languages, preserving both the distribution of error types within a system as well as distribution of each error type over various systems. However, it might not be able to capture very small variations correctly.

The main issue is the low accuracy of inflectional error class due to confusions with mistranslations. For shorter words, actual inflectional errors tend to be tagged as mistranslations, for longer words the other way round. Keeping all that in mind, it is possible to use the error classifier without target language lemmatisation and to extrapolate inflectional and lexical error rates according to the dominant word length in the analysed text.

Our further work will concentrate on combining the $4let$ method with more refined methods which take into account the word length, and also investigating other fixed reduction lengths, e.g. 5 and 6. Comparison with human error classification results as well as manual inspection of problematic words and error confusion types should be carried out as well.

Acknowledgments

This publication has emanated from research supported by QTLEAP project – ECs FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLEAP: Quality Translation by Deep Language Engineering Approaches" and by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289. We are grateful to the reviewers for their valuable feedback.

References

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 22–64, Edinburgh, Scotland, July.

Fraser, Alexander and Daniel Marcu. 2005. ISI's Participation in the Romanian-English Alignment Task. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 91–94, Ann Arbor, Michigan, June.

Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Proceedings of the 8th Language Technologies Conference*, pages 89–94, Ljubljana, Slovenia, October.

Koehn, Philipp and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 03)*, pages 347–354, Budapest, Hungary, April.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, September.

Popović, Maja and Hermann Ney. 2011. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37(4):657–688, December.

Popović, Maja. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, (96):59–68, October.

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 73–80, Leuven, Belgium, May.

Spoustová, Drahomíra "Johanka", Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*, pages 2214–2218, May.

Zeman, Daniel, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: What Is Wrong with My Translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88, October.