PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 96 OCTOBER 2011 59–68

Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output

Maja Popović

DFKI, Language Technology Group

Abstract

We describe Hjerson, a tool for automatic classification of errors in machine translation output. The tool features the detection of five word level error classes: morphological errors, reodering errors, missing words, extra words and lexical errors. As input, the tool requires original full form reference translation(s) and hypothesis along with their corresponding base forms. It is also possible to use additional information on the word level (e.g. POS tags) in order to obtain more details. The tool provides the raw count and the normalised score (error rate) for each error class at the document level and at the sentence level, as well as original reference and hypothesis words labelled with the corresponding error class in text and HTML formats.

1. Motivation

Human error classification and analysis of machine translation output presented in (Vilar et al., 2006) have become widely used in recent years in order to get detailed answers about strengths and weaknesses of a translation system. Another types of human error analysis have also been carried out, e.g. (Farrús et al., 2009) suitable for the Spanish and Catalan languages. However, human error classification is a difficult and time consuming task, and automatic methods are needed.

Hjerson is a tool for automatic error classification which systematically covers the main word level error categories defined in (Vilar et al., 2006): morphological (inflectional) errors, reordering errors, missing words, extra words and lexical errors. It implements the method based on the standard word error rate (WER) combined with the precision and recall based error rates (Popović and Ney, 2007) and it has been

Corresponding author: maja.popovic@dfki.de
Cite as: Maja Popović. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. The Prague Bulletin of Mathematical Linguistics No. 96, 2011, pp. 59–68.
doi: 10.2478/v10108-011-0011-4.

tested on various language pairs and tasks. It is shown that the obtained results have high correlation (between 0.6 and 1.0) with the results obtained by human evaluators (Popović and Burchardt, 2011; Popović and Ney, 2011).

The tool is written in Python, and is available under an open-source licence. We hope that the release of the toolkit will facilitate the error analysis and classification for the researchers, and also stimulate further development of the proposed method.

2. Hjerson Toolkit

2.1. Algorithm

Hjerson implements the edit distance algorithm (Levenshtein, 1966) and identifies actual words contributing to the the standard Word Error Rate (WER) as well as to the recall/precision based Position-independent Error Rates called Reference PER (RPER) and Hypothesis PER (HPER) (Popović and Ney, 2007).

The dynamic programming algorithm for WER enables a simple and straightforward identification of each erroneous word which actually contributes to WER – the WER errors are marked as substitutions, deletions or insertions. The RPER errors are defined as the words in the reference which do not appear in the hypothesis, and analogously, the HPER errors are the words in the hypothesis which do not appear in the reference. Once the WER, RPER and HPER errors have been identified, the base forms for each word are added in order to perform error classification in the following way:

- inflectional error — a word which full form is marked as RPER/HPER error but the base forms are the same.
- reordering error — a word which occurs both in the reference and in the hypothesis thus not contributing to RPER or HPER, but is marked as a WER error.
- missing word — a word which occurs as deletion in WER errors and at the same time occurs as RPER error without sharing the base form with any hypothesis error.
- extra word — a word which occurs as insertion in WER errors and at the same time occurs as HPER error without sharing the base form with any reference error.
- incorrect lexical choice — a word which belongs neither to inflectional errors nor to missing or extra words is considered as lexical error.

Although the method is generally language-independent, availability of base forms for the particular target language is a requisite. If the error classification would be carried out without base forms, the morphological errors could not be detected and the rest of the results would be noisy, which would especially be problematic for morphologically rich(er) languages.

Figure 1 shows the workflow of the procedure. The details about the input and output options are described in following sections.

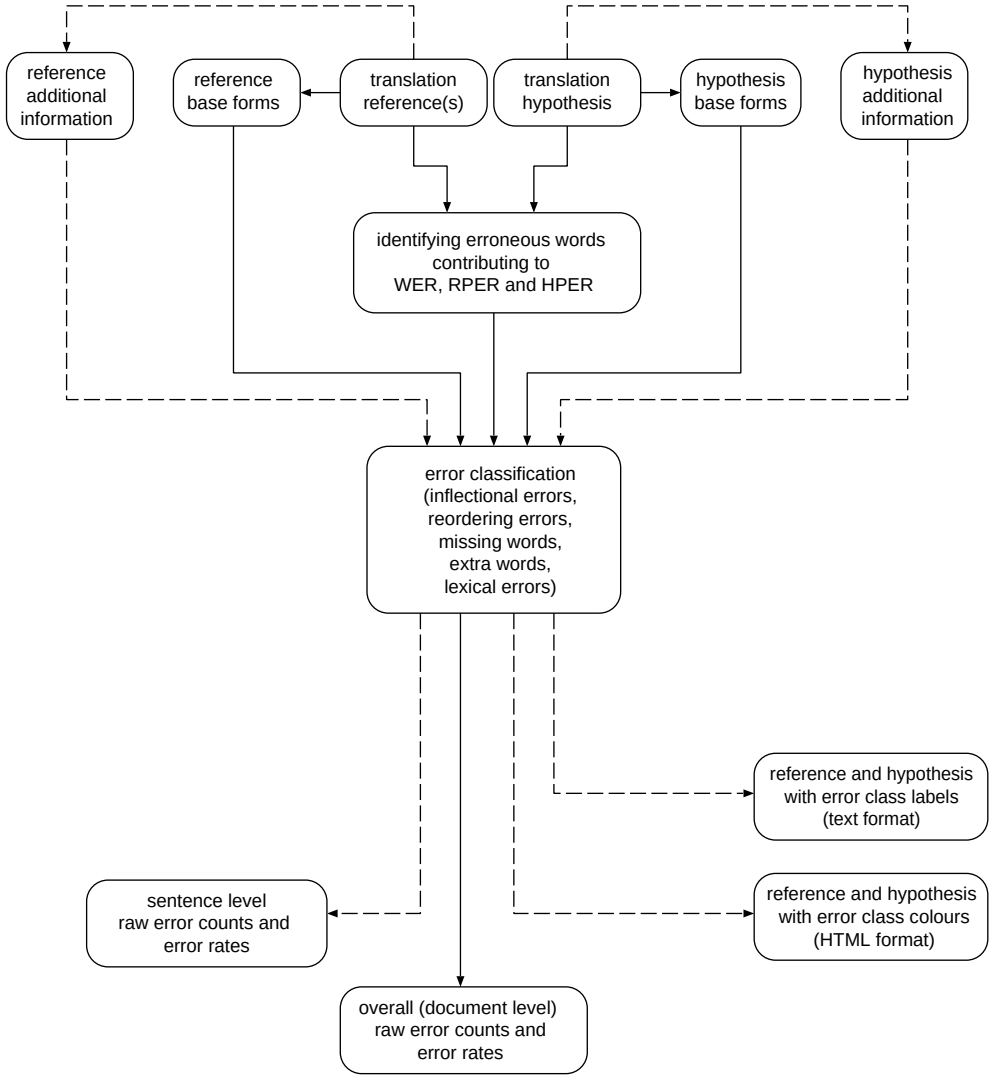

Figure 1. **Workflow of the automatic error classification by Hjerson:** *Continuous lines represent required inputs and default outputs, dashed lines represent optional inputs and outputs.*

2.2. Usage

Hjerson supports the option `-h/--help` which outputs a description of the available command line options.

The input options are:

- `-R, --ref` translation reference
- `-H, --hyp` translation hypothesis
- `-B, --baseref` reference base forms
- `-b, --basehyp` hypothesis base forms
- `-A, --addref` additional reference information
- `-a, --addhyp` additional hypothesis information

Inputs `-R`, `-H`, `-B` and `-b` are required. If any additional information at the word level is available (for example POS tags), it is possible to incorporate it by using options `-A` and/or `-a` in order to obtain more details. The additional information can be provided only for the reference, only for the hypothesis, for both, or not at all.

The required format for all input files is row text containing one sentence per line. In the case of multiple references, all available reference sentences must be separated by the symbol #. For the error classification, the reference sentence with the lowest WER score will be used.

The output options are:

- standard output
  The default output of the tool are the overall (document level) raw error counts and error rates (counts normalised over the reference or hypothesis length) for each of the five error classes:
  - reference and hypothesis inflectional errors (INFER);
  - reference and hypothesis reordering errors (RER);
  - missing words (MISER);
  - extra words (EXTER);
  - reference and hypothesis lexical errors (LEXER).

  For each class, the raw block error counts and block error rates are calculated as well, where block refers to a group of successive words belonging to the same error class. In addition, the values of the initial error rates, i.e. WER, RPER and HPER, are also provided together with their raw error counts.

- `-s, --sent sentence_errors.txt`
  The sentence level raw counts and error rates are written in the given text file `sentence_errors.txt`.

example.hyp	example.ref
This time , the reason for the collapse on Wall Street . The proper functioning of the market and a price .	This time the fall in stocks on Wall Street is responsible for the drop . The proper functioning of the market environment and the decrease in prices .

example.hyp.base	example.ref.base
This time , the reason for the collapse on Wall Street. The proper functioning of the market and a price .	This time the fall in stock on Wall Street be responsible for the drop . The proper functioning of the market environment and the decrease in price .

example.hyp.pos	example.ref.pos
DT NN , DT NN IN DT NN IN NP NP SENT DT JJ NN IN DT NN CC DT NN SENT	DT NN DT NN IN NNS IN NP NP VBZ JJ IN DT NN SENT DT JJ NN IN DT NN NN CC DT NN IN NNS SENT

Table 1. Example of translation hypothesis and its corresponding reference translation.

- `-c, --cats categories.txt`
  This option enables writing original reference and hypothesis words labelled with a corresponding error class in the given text file `categories.txt`. If additional information has been used, it is also contained in this file, which is suitable for potential further processing.

- `-m, --html categories.html`
  The results are writen in the given HTML file `categories.html` where the error classes are visualised by using colours.

An example of input and output files is shown in the next section.

2.3. Example

Table 1 presents an example of translation hypothesis consisting of two sentences and its corresponding reference translation together with their base forms as well as POS tags as additional information.

A program call without additional information:

```
hjerson.py --ref example.ref --hyp example.hyp --baseref example.ref.base
--basehyp example.hyp.base --html example.html --cats example.cats --sent
example.senterrorrates > example.totalerrorrates
```

will produce the following outputs:

- `example.totalerrorrates` – a file containing overall raw counts and error rates:

Wer:	15	53.57			
Rper:	11	39.29			
Hper:	5	22.73			
rINFer:	1	3.57	brINFer:	1	3.57
hINFer:	1	4.55	bhINFer:	1	4.55
rRer:	2	7.14	brRer:	1	3.57
hRer:	2	9.09	bhRer:	1	4.55
MISer:	6	21.43	bMISer:	4	14.29
EXTer:	2	9.09	bEXTer:	2	9.09
rLEXer:	4	14.29	brLEXer:	2	7.14
hLEXer:	2	9.09	bhLEXer:	2	9.09

  where prefixes ”r” and ”h” denote reference and hypothesis, and prefix ”b” denotes blocks.

- `example.senterrorrates` – a file containing raw counts and error rates for each sentence (sentence number is indicated for each error class, for example “1::rRer”).

- `example.html` – a HTML file containing original sentences with visualised error categories: pink (italic) inflectional errors, green (underlined) reordering errors, blue (bold) missing and extra words and red (bold+italic) lexical errors:

  REF: This time the ***fall in stocks*** on Wall Street **is responsible** for the **drop** .
  HYP: This time **,** the **reason** for the ***collapse*** on Wall Street .

  REF: The proper functioning of the market **environment** and **the decrease** ***in*** *prices* .
  HYP: The proper functioning of the market and ***a*** *price* .

- `example.cats` – a text file containing original words labelled with corresponding error category; the label “x” denotes absence of errors, i.e. correct word.

1::ref-err-cats: This~~x time~~x the~~x fall~~lex in~~lex stocks~~lex on~~x Wall~~x Street~~x is~~miss responsible~~miss for~~reord the~~reord drop~~miss .~~x
1::hyp-err-cats: This~~x time~~x ,~~ext the~~x reason~~ext for~~reord the~~reord collapse~~lex on~~x Wall~~x Street~~x .~~x

2::ref-err-cats: The~~x proper~~x functioning~~x of~~x the~~x market~~x environment~~miss and~~x the~~miss decrease~~miss in~~lex prices~~infl .~~x
2::hyp-err-cats: The~~x proper~~x functioning~~x of~~x the~~x market~~x and~~x a~~lex price~~infl .~~x

If POS tags are used as additional information:

```
hjerson.py --ref example.ref --hyp example.hyp --baseref example.ref.base
--basehyp example.hyp.base --addref example.ref.pos --addhyp example.hyp.pos
--html example.html --cats example.cats --sent example.senterrorrates >
example.totalerrorrates
```

the file `example.cats` will contain additional information together with error class labels:

1::ref-err-cats: This#DT~~x time#NN~~x the#DT~~x fall#NN~~lex in#IN~~lex stocks#NNS~~lex on#IN~~x Wall#NP~~x Street#NP~~x is#VBZ~~miss responsible#JJ~~miss for#IN~~reord the#DT~~reord drop#NN~~miss .#SENT~~x
1::hyp-err-cats: This#DT~~x time#NN~~x ,#,~~ext the#DT~~x reason#NN~~ext for#IN~~reord the#DT~~reord collapse#NN~~lex on#IN~~x Wall#NP~~x Street#NP~~x .#SENT~~x

2::ref-err-cats: The#DT~~x proper#JJ~~x functioning#NN~~x of#IN~~x the#DT~~x market#NN~~x environment#NN~~miss and#CC~~x the#DT~~miss decrease#NN~~miss in#IN~~lex prices#NNS~~infl .#SENT~~x
2::hyp-err-cats: The#DT~~x proper#JJ~~x functioning#NN~~x of#IN~~x the#DT~~x market#NN~~x and#CC~~x a#DT~~lex price#NN~~infl .#SENT~~x

The POS tags will also be visible in the HTML file:

REF: This#DT time#NN the#DT ***fall#NN in#IN stocks#NNS*** on#IN Wall#NP Street#NP **is#VBZ responsible#JJ** for#IN the#DT **drop#NN** .#SENT
HYP: This#DT time#NN **,#,** the#DT **reason#NN** for#IN the#DT ***collapse#NN*** on#IN Wall#NP Street#NP .#SENT

REF: The#DT proper#JJ functioning#NN of#IN the#DT market#NN **environment#NN** and#CC **the#DT decrease#NN *in#IN*** *prices#NNS* .#SENT
HYP: The#DT proper#JJ functioning#NN of#IN the#DT market#NN and#CC ***a#DT*** *price#NN* .#SENT

3. Conclusions

We presented Hjerson, a toolkit for automatic error classification which we believe will be of value to the machine translation community. It can be downloaded from `http://www.dfki.de/~mapo02/downloads/hjerson.py`. And for those wondering: Hjerson is a detective solving mysteries (hidden error classes) – he is a recursively fictional character[1] in several books of Agatha Christie.

Acknowledgments

This work has partly been developed within the TARAXÜ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

Bibliography

Farrús, Mireia, Marta R. Costa-jussà, Marc Poh, Adolfo Hernández, and José B. Mariño. Improving a Catalan-Spanish statistical translation system using morphosyntactic knowledge. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT 09)*, pages 52–57, Barcelona, Spain, May 2009.

Levenshtein, Vladimir Iosifovich. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

Popović, Maja and Aljoscha Burchardt. From Human to Automatic Error Classification for Machine Translation Output . In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 265–272, Leuven, Belgium, May 2011.

[1] A fictional character in books written by a fictional character.

Popović, Maja and Hermann Ney. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In *Proceedings of the 2nd ACL 07 Workshop on Statistical Machine Translation (WMT 07)*, pages 48–55, Prague, Czech Republic, June 2007.

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

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pages 697–702, Genoa, Italy, May 2006.

Address for correspondence:
Maja Popović
maja.popovic@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Language Technology Group (LT)
Alt-Moabit 91c
10559 Berlin, Germany