Software Verification and Validation Research Laboratory (SVVRL) of the University of Kentucky: Traceability Challenge 2011: Language Translation

Jane Huffman Hayes, Hakim Sultanov, Wei-Keat Kong, Wenbin Li
University of Kentucky, Lexington, Kentucky
{hayes, hakim.sultanov, wkkong1, wenbin.li}@uky.edu

ABSTRACT
We present the process and methods applied in undertaking the Traceability Challenge in addressing Grand Challenge C-GC1 – Trace recovery. The Information Retrieval methods implemented in REquirementsTRacing On target .NET (RETRO.NET) were applied to the tracing of the eTour and EasyClinic datasets. Our work focused on the nuances of native language (Italian, English). Datasets were augmented with additional terms derived from splitting function and variable names with Camel-Back notation and using the Google Translate API to translate Italian terms into English. Results based on the provided answer set show that the augmented datasets significantly improved recall and precision for one of the datasets.

Categories and Subject Descriptors
D.2.1 [**Requirements/Specifications**]: Tools

General Terms
Measurement

Keywords
Traceability, Translation, Trace Recovery, Challenge

1. INTRODUCTION
Tracing pairs of textual software engineering artifacts is an important yet difficult undertaking. It is important because the resulting traceability matrix (TM) is the basis for regression testing, satisfaction assessment, change impact analysis, etc. Trace recovery is challenging because text artifacts are unstructured; have been written by different authors possibly using different terminology; have been written in different native languages; are at different abstraction levels; contain ambiguous terminology; and can constitute a huge search space.

Automation of tracing is not without its challenges: lack of data sets with true or known answer sets makes validation and comparison of automated tools difficult; information retrieval (IR) techniques suffer from low precision; even more sophisticated techniques such as latent semantic indexing (LSI) or natural language processing (NLP) rules-based approaches require an analyst's approval.

TEFSE'11, May 23, 2011, Waikiki, Honolulu, HI, USA

The Center of Excellence for Software Traceability (COEST) developed a set of Grand Challenges for Traceability (GCT) [1]. This paper addresses challenge C-GC1, trace recovery. Specifically, we focus on traces of use cases to code classes that contain a mixture of Italian and English terms. We ask the question "Can traces between use cases and code classes containing terms in a different language be improved when using a literal translation of terms?" We apply some pre-processing to identify Italian terms and translate them into English. We apply the TF-IDF method to the eTour and EasyClinic datasets and examine the effectiveness of the translated dataset with respect to the provided answer set.

The University of Kentucky Software Verification and Validation Research Laboratory (SVVRL) presents its challenge results here. Our paper is organized as follows. Section 2 explains the process applied by SVVRL in pursuing the challenge. Section 3 describes the technique used to trace two of the Challenge datasets. Section 4 provides results. Finally, Section 5 provides conclusions and future work.

2. CHALLENGE PROCESS
At the start of the challenge, we identified datasets that would be suitable for answering our earlier research question. The eTourand EasyClinicdatasets containedcode classes that had a mixture of Italian and English terms. Although not part of the challenge, we also had access to an Italian version of the EasyClinic dataset. We wanted to see if tracing can be improved with a dataset that is not written in English. These datasets were converted into the format required by our tool REquirementsTRacing On target for .NET (RETRO.NET) [3]. A copy of the original datasets was made and augmented with translated terms. We applied the Term Frequency-Inverse Document Frequency (TF-IDF) method[4] with stopword removal and stemming [5] to the original and augmented datasets. Results of each run were compared to the answer set using standard as well as secondary IR measures.

3. AUGMENTING DATASETS
A parser identified function names and "Camel-Back" notation (where several words are concatenated together) in code classes, splitting them into individual terms before translating them using the Google Translate API [6][1]. Often, such notation provides additional context not picked up by term-based IR techniques, for example, "searchPreferences." We also noticed that some functions still contained Italian words, such as: "OrarioApertura," translated as "Time Opening." Splitting these terms increases the number of terms that could possibly match terms used in the use cases. Translated terms were then compiled into a list that is used

[1]http://code.google.com/apis/language/translate/overview.html

by an automated script to insert translated terms next to the original terms.

4. PRELIMINARY RESULTS

Figure 1 depicts the performance of the TF-IDF technique using the three original (baseline) and three augmented (translated) datasets. The first column of graphs represents the precision-recall curves for the baseline and translated datasets showing the precision-recall drift when relevance scores were filtered from 0 to 0.95. *Recall* is evaluated as the total number of relevant retrieved documents divided by the total number of relevant documents in the whole collection:

$$Recall = \frac{\#of\_relevant\_retrieved}{\#\_relevant\_in\_collection}$$

Precision is evaluated as the total number of relevant retrieved documents divided by the total number of retrieved documents:

$$Precision = \frac{\#of\_relevant\_retrieved}{\#\_retrieved}$$

The second column of graphs represents the Receiver Operating Characteristics (ROC) curves [4] for the baseline and translated datasets. The ROC plots a graph for true positive rate as function of false positive rate (FPR). These graphs plot the FPR over recall, representing how fast false links were removed with each filtering level. The augmented eTour dataset produced better trace results compared to the baseline dataset, consistently beating the baseline at each point on the precision-recall (P-R) curve and the ROC curve. The augmented English EasyClinic showed little difference on the P-R curve and the ROC curve. The augmented Italian EasyClinic dataset, however, showed better results at higher recall levels compared to the baseline. Results were mixed at low to middle recall levels. The Italian EasyClinic ROC curve showed that the augmented dataset performed better than the baseline.

Table 1 shows the secondary measures of MAP [4] and Lag [2] for the three original and augmented datasets. MAP measures "the quality across the recall levels". The higher the MAP, the closer the true links are to the top of the candidate link list. For h_j in a set of textual artifacts $H=\{h_1,...,h_n\}$, a subset of relevant documents $\{d_1,...,d_{m_j}\}$, and $L_{jT} \in L=\{(d,h)|sim(d,h)\}$, a subset of true links ranked by relevance, MAP is evaluated as follows:

$$MAP(H) = \frac{1}{|H|}\sum_{j=1}^{|H|}\frac{1}{m_j}\sum_{k=1}^{m_j} precision(L_{jT})$$

The Wilcoxon Signed-Ranks Test (non-parametric equivalent of the paired t-test) [2] is applied to determine the statistical significance of the difference in MAP and Lag results between the original and augmented datasets. Significant results are bolded in the Table. For the eTour dataset, MAP increased 24% from 0.419 to 0.512 ($p < 0.0001$). MAP did not significantly improve in the case of EasyClinic. Lag for Italian EasyClinic improved significantly from 3.1 to 2.3 ($p < 0.0174$). Lag for Etour also improved significantly from 23.3 to 15.5 ($p < 0.0001$).

[2] http://faculty.vassar.edu/lowry/wilcoxon.html

Table 1. Lag and MAP for all three datasets

Dataset	Lag	MAP
EasyClinic-ENG	3.0	0.738
EasyClinic-ENG-trans	2.7	0.740
EasyClinic-ITA	3.1	0.717
EasyClinic-ITA-trans	**2.3**	0.721
Etour	23.3	0.419
Etour-trans	**15.5**	**0.512**

Table 2, at the end of this paper, shows the specific precision/recall/FPR values for each dataset. Row 2, for example, shows the results for each dataset when filtering out links with relevance scores below 0.05. The eTour dataset had 0.45 recall, 0.26 precision, and 0.08 FPR, while the translated eTour dataset had 0.65 recall, 0.24 precision, and 0.12 FPR. Overall, the recall values for the translated eTour dataset were higher compared to the original dataset at the threshold filter values above 0.1. The translation techniques helped to "expand" terms within the lower level documents, i.e., increased the number of common terms between the two datasets. At the same time, the precision was slightly less for the same threshold values. This can also be explained by retrieving a greater number of documents for inspection, due to the increased number of common terms. For every threshold value, the gains in recall were at least twice as much as the losses in precision.

5. CONCLUSIONS AND FUTURE WORK

We noted a number of interesting items in the eTour dataset. There were quite a few Italian terms that could be extracted from function and variable names in the code classes. The naming convention in the files was consistent, which greatly aided in finding those terms. A literal translation of terms that are possibly Italian could lead to some inaccuracy as well as confusion. Due to lack of knowledge in Italian, we could not verify the accuracy of the Google Translator API. For example, in one of the code classes we saw functions named "ottieniPuntoDiRistoro," which can be parsed and translated as "get point of refreshment." The term "refreshment" in this instance can mean "food," "drink," "rest," or possibly "restaurant." The term "refreshment", however, might not be used in the use case. Perhaps using a thesaurus to map all these terms to "ottieniPuntoDiRistoro" could help us to obtain better recall and precision. Building a thesaurus is still a manual and non-trivial task. One possible direction for future research may be to automatically build a thesaurus when dealing with bi-lingual datasets. Basically, one would build a collection of terms that occur in both collections and then use thesaurus term expansion for common terms, especially for the translated terms. As a result, the search space will be expanded and hopefully reduce the possibility of "getting lost in translation."

There were interesting items in the EasyClinic dataset as well. First, there were a few Italian terms that remained in the English dataset. Second, there was a lack of punctuation in the dataset that made it difficult to trace manually. In addition, words were not organized into completed sentences and seemed to have been extracted from tables. TF-IDF, however, looks at the dataset as a "bag of words" and thus has no issues dealing with the organization of the terms.

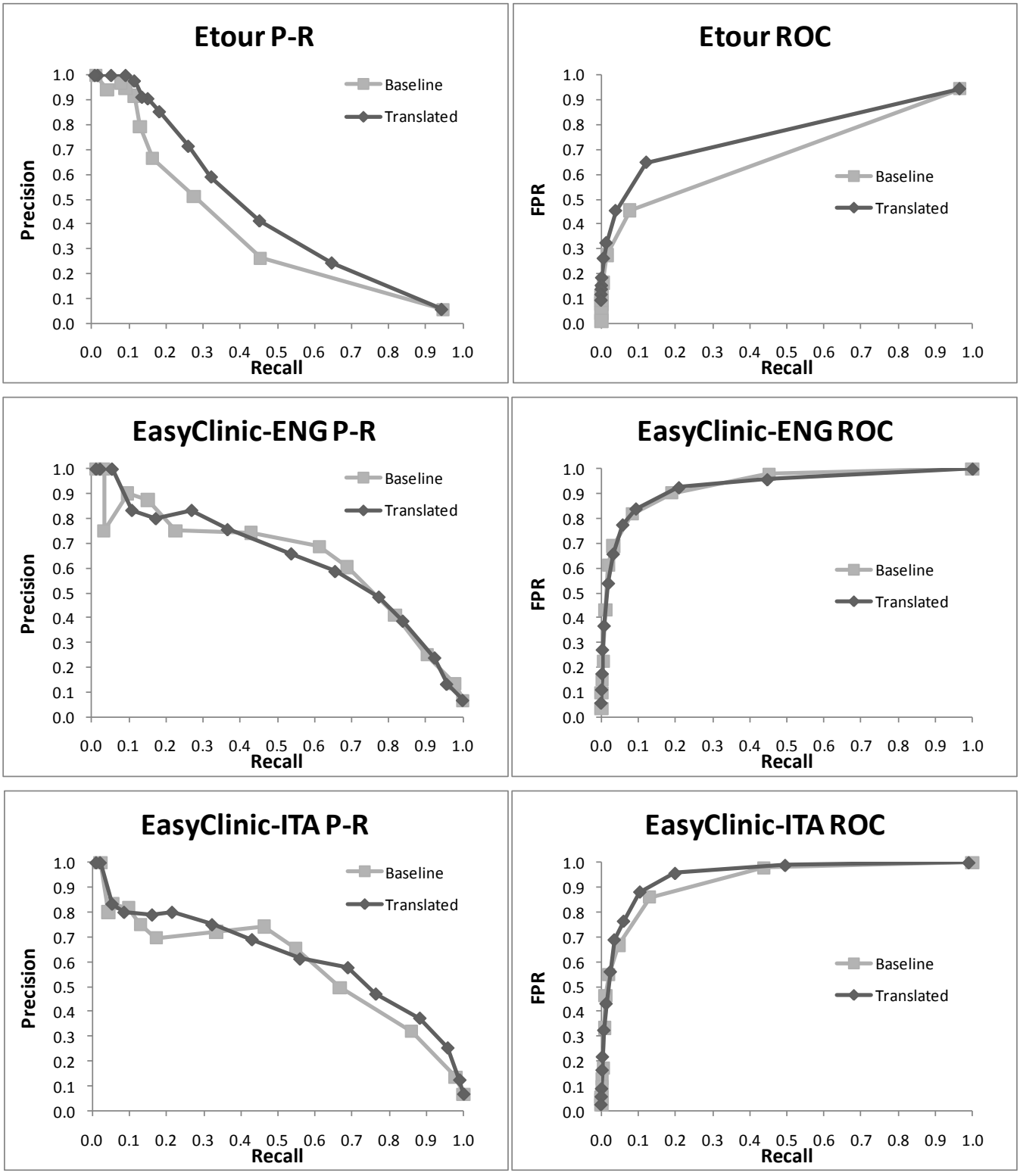

Figure 1. Precision-Recall and ROC graphs for all three datasets

Table 2. Recall, Precision, and FPR for all datasets at each filter level

Filter	EasyClin-ENG			EasyClinic-ENG-trans			EasyClinic-ITA			EasyClinic-ITA-trans			Etour			Etour-trans		
	FPR	Recall	Precision	FPR	Recall	Precision	FPR	Recall	Precision	FPR	Recall	Precision	FPR	Recall	Precision	FPR	Recall	Precision
0.00	1.00	1.00	0.07	1.00	1.00	0.07	1.00	1.00	0.07	0.99	1.00	0.07	0.97	0.95	0.06	0.96	0.94	0.06
0.05	0.45	0.98	0.13	0.45	0.96	0.13	0.44	0.98	0.14	0.50	0.99	0.12	0.08	0.45	0.26	0.12	0.65	0.24
0.10	0.19	0.90	0.25	0.21	0.92	0.24	0.13	0.86	0.32	0.20	0.96	0.25	0.02	0.28	0.51	0.04	0.45	0.41
0.15	0.08	0.82	0.41	0.09	0.84	0.39	0.05	0.67	0.50	0.10	0.88	0.37	0.01	0.16	0.66	0.01	0.32	0.59
0.20	0.03	0.69	0.60	0.06	0.77	0.48	0.02	0.55	0.65	0.06	0.76	0.47	0.00	0.13	0.79	0.01	0.26	0.71
0.25	0.02	0.61	0.69	0.03	0.66	0.59	0.01	0.46	0.74	0.04	0.69	0.58	0.00	0.11	0.92	0.00	0.18	0.85
0.30	0.01	0.43	0.74	0.02	0.54	0.66	0.01	0.33	0.72	0.03	0.56	0.61	0.00	0.09	0.95	0.00	0.15	0.91
0.35	0.01	0.23	0.75	0.01	0.37	0.76	0.01	0.17	0.70	0.01	0.43	0.69	0.00	0.08	0.97	0.00	0.14	0.91
0.40	0.00	0.15	0.88	0.00	0.27	0.83	0.00	0.13	0.75	0.01	0.32	0.75	0.00	0.04	0.94	0.00	0.11	0.98
0.45	0.00	0.10	0.90	0.00	0.17	0.80	0.00	0.10	0.82	0.00	0.22	0.80	0.00	0.01	1.00	0.00	0.09	1.00
0.50	0.00	0.03	0.75	0.00	0.11	0.83	0.00	0.05	0.83	0.00	0.16	0.79					0.05	1.00
0.55	0.00	0.03	1.00	0.00	0.05	1.00	0.00	0.04	0.80	0.00	0.09	0.80					0.02	1.00
0.60		0.01	1.00		0.05	1.00	0.00	0.02	1.00	0.00	0.05	0.83					0.01	1.00
0.65					0.02	1.00				0.00	0.02	1.00						
0.70					0.01	1.00					0.01	1.00						

The EasyClinic dataset had much higher MAP compared to the eTour dataset. Datasets that already have good results from using the TF-IDF method usually do not benefit much from more advanced techniques based on anecdotal evidence. The eTour dataset, having a low MAP, seemed to benefit from the additional term expansion using the "Camel-Back" splitting as well as the literal translations. Future work involves predicting when advanced techniques can be applied to supplement or replace the TF-IDF method.

We note that our results are very similar for the English and ItalianEasyClinic dataset. This is an encouraging sign (though not unexpected) for global application of the literal translation technique, regardless of language. Future work involves locating other software artifacts that are written in other languages and validating the methods described in this paper.

6. ACKNOWLEDGMENTS

This work is funded in part by the National Science Foundation under NSF grants CCF-0811140 (research) and ARRA-MRI-R2 500733SG067 (benchmark development). This work was partially sponsored by NASA under grant NNG05GQ58G.

7. REFERENCES

[1] Cleland-Huang, J., Dekhtyar, A., Hayes, J.H., Antoniol, G.,Berenbach, B., Eyged, A., Ferguson, S., Maletic, J., and Zisman, A.. "Grand Challenges in Traceability," Technical Report COET GCT-06-01-0.9, Center of Excellence for Traceability, September 2006.

[2] Hayes, J.H., Dekhtyar, A., and Sundaram, S.K., "Advancing Candidate Link Generation for Requirements Tracing: The Study of Methods," IEEE Trans. Softw. Eng., vol. 32, 2006, pp. 4-19.

[3] Hayes, J.H., Dekhtyar, A., Sundaram, S.K., Holbrook, E., Vadlamudi, S. and April, A. 2007. REquirementsTRacing On target (RETRO): improving software maintenance through traceability recovery. Innovations in Systems and Software Engineering. 3, (2007), 193-202.

[4] Manning, C.D., Raghavan, P., and Schtze, H., Introduction to Information Retrieval, Cambridge University Press, 2008.

[5] Porter, M.F. 1997. An algorithm for suffix stripping. Readings in information retrieval. Morgan Kaufmann Publishers Inc. 313-316

[6] Google Language API Family, http://code.google.com/apis/language/