Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-132

Bootstrapping of a Multilingual Transliteration Dictionary for European Languages

Mārcis PINNIS[a1]
[a]Tilde, Latvia
University of Latvia, Latvia

Abstract. Transliteration dictionaries are an important resource for the development of machine transliteration systems. The paper describes and analyses a large multilingual transliteration dictionary extracted from probabilistic dictionaries for 24 European languages containing approximately 1.25 million transliterated word pairs. The transliteration dictionary is evaluated: 1) manually for the Latvian-English language pair and 2) automatically within a statistical machine translation based transliteration task for all 23 language pairs.

Keywords. Transliteration, European languages, machine translation

Introduction

Transliteration, which is the process of representing words from one language using the writing system of another language [1][2], is a typical method for the translation of named entities and technical terms [3] (often applying grapheme-to-phoneme and phoneme-to-grapheme transformation rules in the translation process). Creation of a rule-based system can be very time consuming, and therefore an alternative is to build supervised machine learning based methods (e.g., using statistical machine translation technology [4]). However, to build supervised transliteration models that could be integrated in machine translation systems, we require a transliteration dictionary. Although there are multilingual named entity dictionaries (e.g., JRC Names [5], HeiNER [6], and others) available, they are not directly applicable for development of transliteration models, because named entities often contain words which are not transliterated. For example, the organisation name "*European Union*" when translated into Latvian ("*Eiropas Savienība*") contains a transliterated and a translated word.

Therefore, to address the necessity of transliteration dictionaries, this paper presents a method for transliteration dictionary extraction using a bootstrapping process from existing dictionaries, e.g., automatically extracted probabilistic dictionaries [7] or manually created dictionaries containing words in their canonical (or lemma) forms. The paper describes and analyses a large multilingual transliteration dictionary extracted from probabilistic dictionaries for 24 European languages (23 language pairs with English as a source language).[1]

[1] Corresponding Author: Mārcis Pinnis; E-mail: marcis.pinnis@tilde.lv

1. Bootstrapping Method

To create a transliteration dictionary, the author starts with existing Giza++ [8] probabilistic dictionaries extracted from the DGT-TM [9] (for official languages of the European Union) and MultiUN [10] (for English-Russian) parallel corpora. The transliteration dictionaries are bootstrapped from the probabilistic dictionaries in two (or more) steps:

1) In the first step, we apply Romanisation [3] rules to all non-English words. The Romanisation rules have been developed earlier by Pinnis [11] and define one-to-one (e.g., the Greek "*β*" and the Bulgarian "*б*" correspond to the English letter "*b*", etc.), one-to-many (e.g., the Greek "*φ*" corresponds to the English "*th*", the Russian "*ч*" corresponds to the English "*ch*", etc.), and one-to-none (e.g., the Russian letters "*ъ*" and "*ь*" are deleted) correspondences of letters from a non-English alphabet into the English alphabet. Then, we compare the English words to the Romanised words with a string similarity metric based on the Levenshtein distance [12]:

$$Sim(s_1,s_2) = \frac{max\left(len(s_1),len(s_2)\right)\text{-}LevenshteinDistance(s_1,s_2)}{max\left(len(s_1),len(s_2)\right)} \quad (1)$$

   Word pairs exceeding an empirically set threshold of *0.7* are extracted as reciprocal transliterations for the further bootstrapping steps.

2) In the second step (and further steps if necessary), we use the transliterations identified in the previous step to build character-based statistical machine translation (SMT) systems using the Moses SMT toolkit [13]. The SMT systems are used to transliterate entries of the initial dictionary. For the experiments presented in this paper, we use the top five SMT transliterations for each non-English word. New transliteration pairs are identified using the same similarity function from Equation 1.

2. Data Formats

The extracted multilingual transliteration dictionary is stored in an XML document. The dictionary consists of source entries in English (the "*SEntry*" tag in Figure 1). For each source entry, the dictionary provides a list of transliterations in target languages (the "*TEntry*" tags). For each transliteration entry, the dictionary provides the number of the bootstrapping iteration in which the transliteration pair has been identified and the bootstrapping method's confidence score (the Levenshtein distance based similarity).

```
<?xml version="1.0" encoding="utf-16"?>
<TranslitCollection xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                    xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  ...
  <SEntry str="academy">
    ...
    <TEntry lang="et" str="akadeemia" iteration="2" score="0.75" />
    <TEntry lang="lt" str="akademija" iteration="2" score="1" />
    <TEntry lang="lv" str="akadēmija" iteration="2" score="1" />
    ...
  </SEntry>
  ...
</TranslitCollection>
```

Figure 1. Example of the XML format of the multilingual transliteration dictionary

3. Statistics of the Multilingual Transliteration Dictionary

In order to create the multilingual transliteration dictionary, the author performed two bootstrapping iterations. The first bootstrapping iteration produced a total of 598,807 transliteration pairs for 82,454 English words across all 23 language pairs. The second iteration resulted in 1,246,908 transliteration pairs for 104,803 English words.

The quantitative results for English-Latvian (see Table 1) show a significant increase in new transliteration pairs extracted in the second bootstrapping iteration. The increase can be explained by the SMT-based transliteration method's ability to deal with inflectional characteristics of different languages. That is, the SMT translation model learns from parallel data (transliteration equivalents identified in previous steps) to translate language specific word prefixes and suffixes from one language into another. As the rule-based method is not capable of performing such language specific transformations, it cannot identify many good transliteration equivalents.

Table 1 also shows that for English-Latvian, the first two out of five total iterations allow acquiring approximately 97% of all extracted English words. Because the initial dictionaries are exhaustive resources (i.e., they contain a fixed number of entries out of which only a certain amount are potential transliterations) and the first two iterations are able to identify the majority of transliteration equivalents, all further iterations are less productive. The 97% comprise approximately 20% of all 134,146 unique English words present in the initial probabilistic dictionary. Taking into account that English and Latvian are not closely related languages, this is a relatively large number.

As a result, only the first two bootstrapping iterations were performed for the multilingual transliteration dictionary. The statistics of the dictionary for all 23 language pairs with English as the source language are given in Table 2. The extracted pair count for Croatian-English is lower due to the smaller size of the initial probabilistic dictionary.

Table 1. Statistics of new English-Latvian transliteration pairs identified in five bootstrapping iterations.

Iteration	New pairs	% increase	New English words	% increase
1	30,879	-	15,598	-
2	41,347	134%	11,992	77%
3	1,704	2%	500	2%
4	469	1%	125	0%
5	961	1%	255	1%
Total	**72,226**		**28,470**	

Table 2. Statistics of the multilingual transliteration dictionary after merging first and second iteration data.

Target language	Unique English words	Transliteration pairs	Target language	Unique English words	Transliteration pairs
BG	17,567	37,901	LT	25,258	66,243
CS	28,366	58,931	LV	27,590	72,186
DA	27,321	51,383	MT	21,217	62,428
DE	23,862	41,560	NL	23,673	36,741
EL	15,513	31,273	PL	29,723	62,313
ES	35,030	64,480	PT	37,666	67,473
ET	22,188	48,113	RO	27,295	58,531
FI	18,180	33,860	RU	30,835	71,482
FR	33,367	59,390	SK	31,536	77,607
HR	7,368	14,965	SL	30,364	66,365
HU	26,942	53,664	SV	28,692	53,676
IT	31,147	56,343			

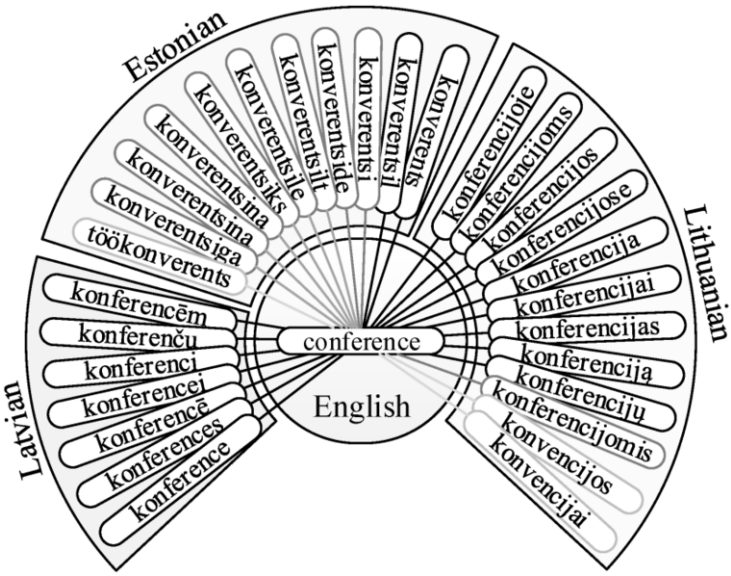

Figure 2. Transliterations of the English word "*conference*" in Estonian, Latvian, and Lithuanian identified in the Giza++ dictionaries extracted from the DGT-TM corpus

A visual example of an entry in the transliteration dictionary for the Baltic languages is given in Figure 2. The light grey to black connectors between English and the target languages indicate low (grey) to high (black) confidence scores assigned to the transliteration pairs by the bootstrapping method.

4. Evaluation

The evaluation of the multilingual transliteration dictionary consists of: 1) manual evaluation for the English-Latvian language pair and 2) automatic evaluation of the transliteration dictionary in an SMT-based transliteration task for 23 language pairs.

4.1. Manual Evaluation

Manual evaluation of the multilingual transliteration dictionary has been performed for the English-Latvian language pair. The author executed a total of five bootstrapping iterations and extracted only newly identified transliteration pairs from each iteration (the quantitative statistics are given in Table 1). Further, 100 transliteration pairs were randomly selected from the newly extracted transliteration pairs for manual evaluation. A transliteration pair in the manual evaluation is considered correct if:

1) The pair consists of words that are reciprocal translations.
2) The pair qualifies to be a transliteration pair. That is, whether we can acquire from the source word the target word (and vice versa) by performing alphabet specific letter transformations (e.g., the Latvian "*č*" can correspond to the English "*ch*", the Greek "*ρ*" can correspond to the English "*r*", etc.) and language specific prefix and suffix transformations (e.g., the English suffix "*ation*" may correspond to the Latvian "*ācija*", Italian "*azione*", the Bulgarian "*ация*", and other suffixes also in many different inflected forms).

The evaluation results are given in Figure 3. The results show that the precision of the transliteration dictionary for English-Latvian is over 90% after the first bootstrapping iteration. Taking into account that the initial probabilistic dictionaries are of very low quality [7], this is a very good result. The figure also shows that the precision of the newly extracted transliteration pairs decreases with each new bootstrapping iteration. Although this was to be expected, the thresholds for different bootstrapping iterations could be differentiated in order to achieve a stable precision of over 90%.

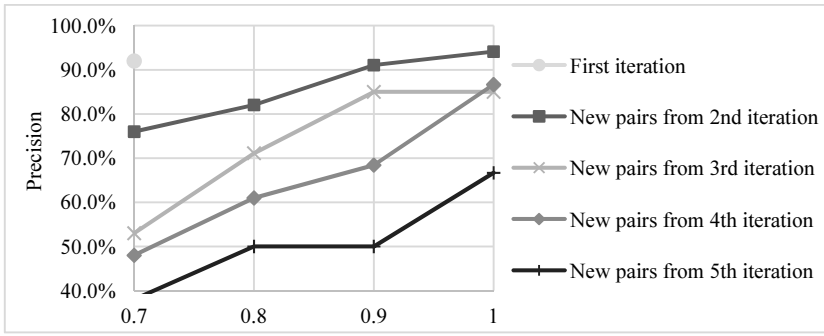

Figure 3. Manual evaluation results for 100 randomly selected transliteration pairs from the English-Latvian transliteration dictionary from different bootstrapping iterations

4.2. Automatic Evaluation in an SMT-based Transliteration Task

Transliteration dictionaries have shown to be beneficial when integrated into SMT systems [4]. However, they are also used for development of machine transliteration systems [3] (e.g., character-based SMT [14]). Recent research has shown that such systems can be used for cross-lingual term alignment in comparable corpora [11]. This paper evaluates the extracted dictionary in an SMT-based transliteration task.

After the second bootstrapping iteration, the source-to-English transliteration data was randomly split in 10 data folds. In each data fold, eight parts were used for training,

one – for tuning, and one – for evaluation. Then, 10-fold cross validation was performed by measuring character level SMT quality using BLEU [15] and NIST [16]. The results are given in Table 3. The results are shown with a 99% confidence interval.

Table 3. Character level 10-fold cross-validation results for character-based SMT transliteration.

Language pair	NIST	BLEU	Language pair	NIST	BLEU
BG-EN	11.48±0.04	90.11±0.23	LT-EN	11.71±0.02	89.49±0.15
CS-EN	12.07±0.03	90.46±0.14	LV-EN	11.87±0.03	89.78±0.22
DA-EN	11.92±0.04	90.37±0.17	MT-EN	11.63±0.04	90.35±0.21
DE-EN	11.89±0.02	90.30±0.17	NL-EN	11.68±0.07	89.42±0.29
EL-EN	10.94±0.04	85.29±0.25	PL-EN	11.96±0.02	89.85±0.18
ES-EN	11.84±0.05	88.20±0.31	PT-EN	11.99±0.05	88.83±0.23
ET-EN	12.13±0.03	91.93±0.20	RO-EN	11.67±0.03	88.67±0.13
FI-EN	12.10±0.05	92.54±0.47	RU-EN	11.23±0.04	83.27±0.18
FR-EN	11.99±0.06	88.39±0.27	SK-EN	12.15±0.05	90.84±0.18
HR-EN	10.53±0.07	87.60±0.31	SL-EN	12.02±0.03	89.71±0.12
HU-EN	12.17±0.02	91.10±0.13	SV-EN	11.92±0.02	89.91±0.14
IT-EN	11.51±0.05	86.78±0.28			

Depending on usage scenarios, an SMT system can be asked to produce one (e.g., for integration of transliteration in machine translation) or many (e.g., for cross-lingual term alignment) transliteration equivalents. Figure 4 shows the precision for up to the top ten SMT generated transliteration equivalents for Baltic languages (results for other language pairs are given in Table 4) when transliterated into English. Because of different inflectional forms in transliteration pairs (e.g., singular vs. plural forms, verbs in different tenses, etc.), the results show a significant increase in precision for the top two to top four transliteration equivalents over the results of the top one.

Another reason for the lower precision for the top one transliteration is the ambiguity of different character sequence transformations, which cannot be predicted by analysis of the surrounding context (letters to the left and to the right). For instance, the differences between writing paradigms in American English and British English allow the Latvian "*organizācija*" to be transliterated as "*organization*" or "*organisation*". Another ambiguous (or non-predictive) example is, for instance, the Latvian "*Kuba*" transliterated in English. It can be either the country "*Cuba*" or a three-dimensional figure "*Cube*". Obviously, the top one transliteration will not always be the expected transliteration because of such ambiguities. A list of the most frequent top

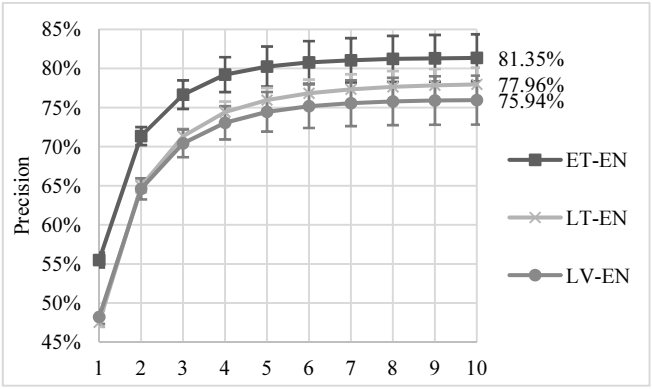

Figure 4. 10-fold cross-validation results for the top *N* SMT transliteration equivalents for Baltic Languages

Table 4. 10-fold cross-validation results for the top 1, top 5, and top 10 SMT transliteration equivalents.

Language pair	Top 1	Top 5	Top 10	Language pair	Top 1	Top 5	Top 10
BG-EN	51.36±0.6%	78.20±2.2%	79.87±2.8%	LT-EN	47.52±0.6%	75.94±1.6%	77.96±2.1%
CS-EN	49.93±0.7%	75.15±1.9%	76.13±2.2%	LV-EN	48.21±0.9%	74.45±2.5%	75.94±3.1%
DA-EN	47.37±0.8%	74.94±1.2%	76.38±1.3%	MT-EN	53.68±0.9%	75.95±3.2%	76.94±3.6%
DE-EN	46.01±0.9%	77.02±2.1%	79.12±2.9%	NL-EN	45.31±1.0%	63.09±2.2%	63.67±2.3%
EL-EN	41.89±0.8%	66.10±1.8%	68.06±2.1%	PL-EN	47.45±0.5%	75.81±2.0%	77.57±2.6%
ES-EN	43.94±0.6%	66.42±1.5%	67.37±1.8%	PT-EN	46.33±0.8%	66.95±2.7%	67.69±3.1%
ET-EN	55.49±1.0%	80.24±2.6%	81.35±3.0%	RO-EN	44.48±0.6%	73.88±1.9%	75.67±2.4%
FI-EN	59.89±1.3%	81.70±1.6%	82.59±1.7%	RU-EN	37.95±0.5%	61.68±1.5%	63.79±1.9%
FR-EN	42.29±1.0%	63.71±3.3%	64.42±3.6%	SK-EN	51.30±0.6%	78.08±1.4%	79.63±2.0%
HR-EN	43.12±1.5%	66.02±4.6%	68.38±5.7%	SL-EN	48.17±0.5%	76.91±2.5%	78.77±3.1%
HU-EN	51.94±0.8%	76.50±2.7%	77.37±3.1%	SV-EN	46.64±0.8%	75.55±1.4%	77.37±1.8%
IT-EN	37.71±0.9%	68.34±2.6%	71.44±3.6%				

Table 5. 15 most frequent character level errors for the Latvian-English SMT-based transliteration system (In the table: Insertions – *Ins*, Deletions – *Del*, Substitutions – *Sub*).

No.	Error	% of all	Latvian (in different inflected forms)	English	
				Expected	Generated
1	Ins/Del *s*	19.79%	zonā	zone**[s]**	zone
			organismus	organism	organism**[s]**
2	Ins/Del *e*	6.42%	krese	cress	cress**[e]**
			validēt	validat**[e]**	validat
3	Ins/Del *a*	3.82%	komponentā	component	component**[a]**
			memorandu	memorand**[a]**	memorand
4	Ins/Del -	3.39%	kvazistatiskas	quasi**[-]**static	quasistatic
			subklīniskas	subclinical	sub**[-]**clinical
5	Ins/Del *al*	3.29%	stratēģiskai	strategic	strategic**[al]**
			teorētiskām	theoretic**[al]**	theoretic
6	Sub *z* ↔ *s*	3.27%	realizējis	reali**[z]**ed	reali**[s]**ed
			organizē	organi**[s]**e	organi**[z]**e
7	Ins/Del *o*	2.66%	luksemburga	luxemburg	luxemb**[o]**urg
			fosforu	phosphor**[o]**us	phosphorus
8	Ins/Del *d*	2.38%	koncentrētos	concentrate**[d]**	concentrate
			neitralizētu	neutralise	neutralise**[d]**
9	Ins/Del *h*	2.37%	homeopātiskas	homeopat**[h]**ic	homeopatic
			metrīta	metritis	met**[h]**ritis
10	Sub *i* ↔ *y*	2.01%	iridovīrusa	[i]ridovirus	**[y]**ridovirus
			elektrolīts	electrol**[y]**te	electrol**[i]**te

one transliteration errors for Latvian-English is given in Table 5. Note that the table shows also ambiguous examples, which are not actual errors, e.g., singular vs. plural forms, different verb tenses, etc.

For the Latvian-English transliteration direction, the SMT-based transliteration quality for systems trained on data from the first and second bootstrapping iterations was also analysed. Although the manual evaluation results show that the overall quality of the data after the second iteration is lower, the SMT evaluation shows that this data allows achieving higher word level precision. The results show that the SMT system is able to build a more generalised translation model by using more data.

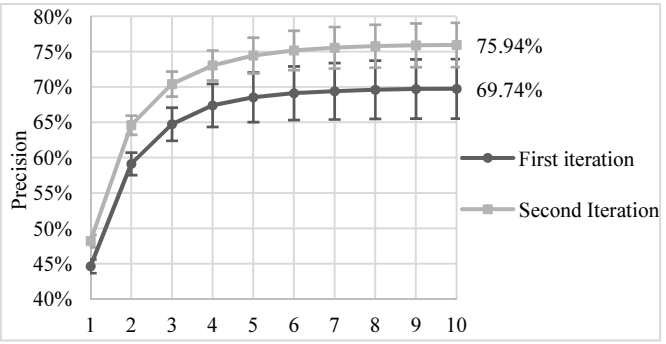

Figure 5. 10-fold cross-validation results for the top *N* SMT generated transliteration equivalents. The chart compares Latvian-English SMT-based transliteration systems trained on the transliteration dictionaries from the first and second bootstrapping iteration. The error bars indicate a 99% confidence interval.

5. Conclusion

In this paper, the author presented a bootstrapping method for the creation of a multilingual transliteration dictionary from existing probabilistic dictionaries. The multilingual transliteration dictionary generated by the author using probabilistic dictionaries extracted from the DGT-TM parallel corpus and the MultiUN parallel corpus covers 24 languages and contains a total of 1,246,908 transliteration pairs. The evaluation has shown that the transliteration dictionary can be effectively applied in SMT-based transliteration tasks. The dictionary and the tools for creation of the dictionary are freely downloadable at https://github.com/pmarcis/dict-filtering.

6. Acknowledgements

This work has been supported by the European Social Fund within the project «*Support for Doctoral Studies at University of Latvia*». The research leading to these results has received funding from the research project "2.6. Multilingual Machine Translation" of EU Structural funds, contract nr. L-KC-11-0003 signed between the ICT Competence Centre (www.itkc.lv) and the Investment and Development Agency of Latvia.

References

[1] Arbabi, M., Fischthal, S. M., Cheng, V. C., & Bart, E. (1994). Algorithms for Arabic Name Transliteration. IBM Journal of Research and Development, 38(2), 183–194.

[2] Pouliquen, B., Steinberger, R., Ignat, C., Temnikova, I., Widiger, A., Zaghouani, W., & Zizka, J. (2005). Multilingual person name recognition and transliteration. Journal CORELA - Cognition, Representation, Language.

[3] Knight, K., & Graehl, J. (1997). Machine transliteration. EACL 1997, 128–135.

[4] Kirschenbaum, A., & Wintner, S. (2010). A General Method for Creating a Bilingual Transliteration Dictionary. In Proceedings of the seventh international conference on Language Resources and Evaluation (LREC-2010), 273–276.

[5] Steinberger, R., & Pouliquen, B. (2011). JRC-Names: A freely available, highly multilingual named entity resource. RANLP 2011, 104–110.

[6] Wentland, W., Knopp, J., Silberer, C., & Hartung, M. (2008). Building a Multilingual Lexical Resource for Named Entity Disambiguation , Translation and Transliteration. LREC 2008.
[7] Aker, A., Pinnis, M., Paramita, M. L., & Gaizauskas, R. (2014). Bilingual dictionaries for all EU languages. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA).
[8] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computa-tional Linguistics, 29, 19–51.
[9] Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. LREC 2012, 454–459.
[10] Eisele, A., & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. LREC 2010, 2868–2872.
[11] Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP 2013), 562–570, Hissar, Bulgaria.
[12] Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10: 707–10.
[13] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation, ACL 2007.
[14] Finch, A., & Sumita, E. (2008). Phrase-based machine transliteration. In Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST), 13–18.
[15] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. In Proceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 2002, 311–318.
[16] Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research (HLT 2002), 2002, 138–145, San Diego, USA.