electronics

MDPI

Article

Hybrid Translation with Classification: Revisiting Rule-Based and Neural Machine Translation

Jin-Xia Huang [1,*], Kyung-Soon Lee [2,*] and Young-Kil Kim [1]

[1] Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon 34129, Korea; kimyk@etri.re.kr (Y.-K.K.)

[2] Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 561756, Korea

* Correspondence: hgh@etri.re.kr (J.-X.H.); selfsolee@jbnu.ac.kr (K.-S.L.)

Received: 7 December 2019; Accepted: 20 January 2020; Published: 21 January 2020

Abstract: This paper proposes a hybrid machine-translation system that combines neural machine translation with well-developed rule-based machine translation to utilize the stability of the latter to compensate for the inadequacy of neural machine translation in rare-resource domains. A classifier is introduced to predict which translation from the two systems is more reliable. We explore a set of features that reflect the reliability of translation and its process, and training data is automatically expanded with a small, human-labeled dataset to solve the insufficient-data problem. A series of experiments shows that the hybrid system's translation accuracy is improved, especially in out-of-domain translations, and classification accuracy is greatly improved when using the proposed features and the automatically constructed training set. A comparison between feature- and text-based classification is also performed, and the results show that the feature-based model achieves better classification accuracy, even when compared to neural network text classifiers.

Keywords: hybrid machine translation; neural machine translation; rule-based machine translation; feature-based classification

1. Introduction

Over the past few years, automated translation performance has improved dramatically with the development of neural machine translation (NMT). In the past decades, we said that rule-based machine translation (RBMT) had a high meaning-transmission accuracy, and statistical-based machine translation (SMT) had excellent fluency. However, NMT has excellent quality in both aspects when there is a large-scale, bilingual parallel corpus.

In practice, it is often difficult to acquire such large-scale parallel corpora, except with some special language pairs or for specific companies. Additionally, most of the special-domain corpora are limited. Considering the weakness of NMT on low-resource and out-of-vocabulary issues, some researchers proposed using a hybrid approach of aiding NMT with SMT [1]. For teams that have already developed an RBMT system with a stable translation quality, utilizing their existing RBMT with NMT when translating a specific domain can be an attractive solution. To test this assumption, we performed a preliminary evaluation on the Korean-to-Chinese translation performance of RBMT, SMT, and NMT in several different domains, including news, Twitter, and spoken language. The spoken dataset with 1012 bilingual sentence pairs consists of daily conversations, travel sentences, and clean sentences selected from the log of the mobile translation service. The other spoken dataset and the single-word dataset are composed with complete sentences and single words, respectively, are randomly selected from the same translation service log. News and Twitter are randomly collected from the day's news and hot issue-related trending topics on Twitter. Except for the first

spoken dataset with bilingual sentence pairs, other datasets do not have translation references—we did not translate the data manually to perform BLEU evaluation [2]. Instead, our translators evaluated the MT results directly, because it would be more intuitive and faster for small datasets. The average length in Table 1 represents the average number of word segments (spacing unit in Korean) per sentence in difference datasets. GenieTalk (HancomInterfree, Seoul, Korea), Moses [3], and OpenNMT [4] are adopted for RBMT, SMT, and NMT, respectively, and 2.87 million bilingual sentence pairs in the spoken-language domain are used to train the SMT and NMT systems. BLEU and translation accuracy [5,6] are adopted for large- and small-scale assessments, respectively.

Table 1 shows that both NMT and SMT outperform RBMT for in-domain translation ("Spoken"). However, the translation quality drops significantly for out-of-domain sentences. This tendency is more severe in the case of NMT, showing that NMT is more vulnerable in low-resource domains ("News" and "Twitter") and is more sensitive to noisy data ("Twitter"). NMT and SMT are also weaker when performing translation in the absence of context information ("Single word"), while RBMT shows better quality in these domains.

Table 1. Translation quality comparison of rule-based machine translation (RBMT), statistical-based machine translation (SMT), and neural machine translation (NMT).

Domain	Evaluation matrix	Sentence count	Average length	RBMT	SMT	NMT
Spoken	BLEU	1012	3.70	0.2570	0.3671	0.4273
Spoken	Accuracy	30	2.90	84.17%	89.58%	90.42%
News	Accuracy	20	17.05	68.13%	46.88%	42.50%
Twitter	Accuracy	20	13.70	46.25%	46.25%	35.00%
Single word	Accuracy	10	1.00	96.25%	80.00%	82.50%

Starting from the above observations, this paper proposes combining RBMT with NMT using a feature-based classifier to select the best translation from the two models; this is a novel approach for hybrid machine translation.

The contribution of this paper resides in several aspects:

- To the best of our knowledge, this paper is the first to combine NMT with RBMT results in domains where language resources are scarce, and we have achieved good experimental results.
- In this paper, we propose a set of features, including pattern matching features and rule-based features, to reflect the reliability of the knowledge used in the RBMT transfer process. This results in better performance in quality prediction than surface information and statistical information features, which have been widely used in previous research.
- Since neural networks perform text classification well, we compare feature-based classification with text-based classification, and our results show that feature-based classification has better performance, which means explicit knowledge that expresses in-depth information is still helpful, even for neural networks.
- To construct a training corpus for the classification of hybrid translation, which is very costly, our study built a small hybrid training corpus manually and then automatically expanded the training corpus with the proposed classifier. As a result, classification performance was greatly improved.

The rest of this paper is organized as follows: Section 2 reviews related work on hybrid machine translation (MT) systems. Section 3 presents our hybrid system's architecture and the specific MT systems adopted in this paper. Section 4 presents our proposed features for hybrid machine translation. Section 5 describes the experimental settings for the tests and discusses the results from different experiments. We present our conclusions and discuss future research in Section 6.

2. Related Works

The first effort toward a hybrid machine translation system adopted SMT as a post-editor of RBMT. In this model, a rule-based system first translated the inputs, and then an SMT system

corrected the rule-based translations and gave the final output [7–11]. The training corpus for the post-editing system consisted of a "bilingual" corpus, where the source side was the rule-based translation, and the target side was the human translation in the target language. This architecture aimed to improve lexical fluency without losing RBMT's structural accuracy. The experiments showed that the purpose was partially achieved: automatic post-editing (APE) outperformed both RBMT and SMT [9,10]. On the other hand, APE for RBMT degraded the performance on some grammatical issues, including tense, gender, and number [9]. It was particularly limited in long-distance reordering for language pairs that have significant grammatical differences; analysis errors could further exacerbate this because of structural errors introduced during rule-based translation [11].

Other researchers proposed an "NMT→SMT" framework, which combined NMT with SMT in a cascading architecture [1]. NMT was adopted as a pre-translator, and SMT, which was tuned by a pre-translated corpus, was adopted as a target–target translator. As a result, the performance was significantly improved in Japanese-to-English and Chinese-to-English translations.

APE has developed to a professional level where human post-edits of the automatic machine translation are required instead of independent reference translations [12–14], because learning from human post-edits is more effective for post-editing [15–16]. However, developing a human-edit corpus is time-consuming and costly, so it cannot be performed for all language pairs. There is not a clear performance improvement when NMT results are post-edited using neural networks, unlike when SMT results are post-edited using neural networks [17].

Reference [18] reported that under the same conditions, selecting the best result from several independent automatic translations was better than APE. This hybrid approach adopted several independent translation systems and picked the best result using either quality estimation or a classification approach. Quality estimation predicts the translation qualities using the language model, word alignment information, and other linguistic information; then, it ranks the quality scores of multiple translations and outputs the top-ranked result as the final translation [19,20]. In addition to hybrid MT, quality estimation also has other applications, such as providing information about whether the translation is reliable, highlighting the segments that need to be revised and estimating the required post-editing effort [21–23].

Feature extraction is normally one of the main issues of quality estimation and classification-based hybrid translation methods. Features can be separated into black-box and glass-box types: black-box features can be extracted without knowing the translation process, include the length of source and target sentences, the n-gram frequency, and the language model (LM) probability of the source or target segments. Meanwhile, glass-box features depend on some aspect of the translation process, such as the global scores of the SMT system and the number of distinct hypotheses in the n-best list of SMT [22,23]. Most of the existing research has either focused on black-box features or the glass-box features from SMT, including language model perplexity, translation ambiguity, phrase-table probabilities, and translation token length [5,18,24–26]. However, a classification approach is considered more proper for a hybrid RBMT/SMT system, because most of the quality estimation research tend to evaluate translation quality using language models and word alignment information; thus, it tends to overvalue the results of SMT [5,18].

Another issue with the classification approach involves the training corpus for hybrid classification [5,24,25,27,28]. The training corpus is composed of extracted features and labels indicating the "better" and "best" translations. Human translators should be involved in building the labeled training corpus; therefore, building such a corpus is a time-consuming and costly process. In previous researches, given the bilingual corpus and the outputs of the two MT systems, the labels are determined based on the quality evaluation metrics such as BLEU. However, this approach has a limit in hybridizing SMT with RBMT, because such measures may cause a biased preference for language model-based systems [5]. Reference [5] proposed an auto-labeling method for RBMT and SMT hybridization that only evaluated SMT translations: the labels would indicate that the SMT results were better if their BLEU metrics were high enough, and the labels would indicate that the RBMT results were better if the SMT results' BLEU metrics were low enough. This

approach avoided the difficulty of securing a large-scale training corpus, but it assumed that an SMT translation was better than an RBMT translation based on the SMT translation's BLEU score; the quality of the RBMT translation was not considered. In practice, sentences that are translated well by SMT are relatively short and contain high-frequency expressions; these are also translated well by RBMT. Sentences that receive low scores with SMT also show relatively inferior performance with RBMT.

Other researchers adopted only one of these systems, either rule-based or statistical, as the basis, letting other systems produce sub-phrases to enrich the translation knowledge. Reference [29] adopted an SMT decoder with a combined phrase table (produced by SMT and several rule-based systems) to perform the final translation. Others [28,30–32] used a rule-based system to produce a skeleton of the translation and then decided whether the sub-phrases produced by SMT could be substituted for portions of the original output. This architecture required the two systems to be closely integrated, making implementation more difficult.

3. System Architecture

This paper aims to offset the disadvantages of NMT, which shows low performance in low-resource domains, by using RBMT. As shown in Figure 1, a classification approach is adopted to select the best of the RBMT and NMT translations.

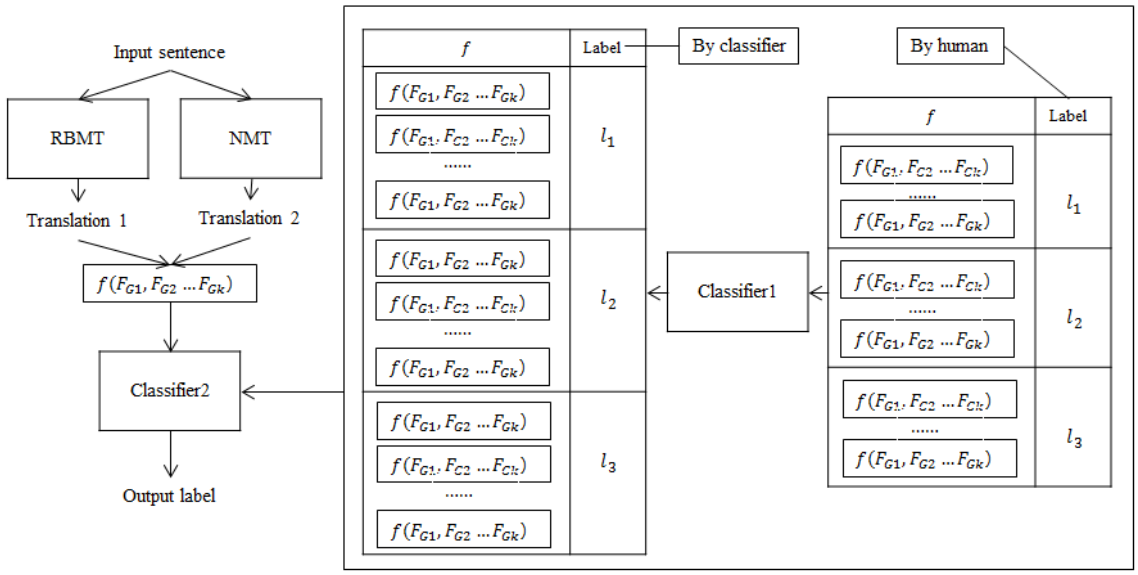

Figure 1. Hybrid system architecture. Classifier 1 is trained on the human-labeled dataset to automatically build a larger dataset. The automatically classified dataset is used with the human-labeled data to train the Classifier 2 for hybrid translation.

Since the performance of RBMT is much lower than that of NMT in general cases, classification accuracy becomes more important to prevent the hybridized results from being lower than those of NMT, which is the baseline. To ensure accurate classification, we thoroughly explored glass-box features that reflect the confidence of RBMT, and we expanded the training corpus automatically using a small-capacity, labeled corpus constructed by human developers, which is a decidedly simple self-training approach.

3.1. Rich Knowledge-Based RBMT

The RBMT system adopted in this study is a machine translation system based on rich knowledge [33]. The main knowledge sources used for the transfer procedure include a bilingual dictionary (backed by statistical and contextual information) and large-scale transfer patterns (at the phrase and sentence level). The phrase-level patterns, which play the most important role in the

transfer procedure, describe dependencies and syntactic relations with word orders in the source and target languages. The arguments of the patterns can be at the part-of-speech (POS) level, semantic level, or lexicon level. The following are examples of Korean-to-English transfer patterns; for readability, Korean characters have been replaced with phonetic spellings:

- (N1:dobj) V2 → V2 (N1:dobj)
- (N1[sem=location]:modi+ro/P2) V3 → V3 (N1:dobj)
- (N1[sem=tool]:dobj) gochi/V2 → fix/V2 (N1:dobj)
- (N1[sem=thought]:dobj) gochi/V2 → change/V2 (N1:dobj)
- (hwajang/N1:dobj) gochi/V2 → refresh/V2 (makeup/N1:dobj)

The source-language arguments are used as constraints in pattern matching. A pattern with more arguments, or with a higher proportion of lexical- and semantic-level arguments, tends to be more informative. We use weights to express the amount of information in the source-language portion of the patterns. The weight of the lexical-level argument is higher than the semantic argument, and the weight of the semantic argument is higher than the POS argument. In terms of POS, verbs and auxiliary words weigh more than adjectives, while adjectives weigh more than nouns. Adjectives weigh more than nouns, because Korean adjectives have various forms of use, which cause the ambiguity in both analysis and transfer. For example, the Korean phrase "yeppeuge ha (keep sth pretty)" can be "yeppeu/J+geha/X", or "yeppeuge/A ha/V". In the previous analysis, the combination of adjectives and auxiliary words increased the difficulty of translation. Patterns that obtain higher matching weights gain higher priority in pattern matching. The quality of the translation strongly depends on the patterns used in the transfer process.

The system supports multilingual translations from and to Korean, so the patterns share the same form in different language pairs. The following Korean-to-Chinese transfer patterns share the same Korean parts with the above Korean-to-English patterns, while some of them even share the same target-language parts (if there is no lexicon argument involved):

- (N1:dobj) V2 → V2 (N1:dobj)
- (N1[sem=location]:modi+ro/P2) V3 → V3 (N1:dobj)
- (N1[sem=tool]:dobj) gochi/V2 → xiu1li3/V2 (N1:dobj)
- (N1[sem=thought]:dobj) gochi/V2 → gai3/V2 (N1:dobj)
- (hwajang/N1:dobj) gochi/V2 → bu3zhuang1/V2_1

3.2. Neural Machine Translation

The NMT system OpenNMT [4], which uses an attention-based, bi-directional recurrent neural network, is adopted in this paper. The corpus used for training covers travel, shopping, and diary domains as Korean–Chinese language pairs. As described in Section 1, there are 2.87 million pairs. The sentences in both languages are segmented with byte-pair encoding [34] to minimize the rare-word problem. The dictionaries include 10,000 tokens in Korean and 17,400 tokens in Chinese. The sizes of the dictionaries are so small that most tokens are character-level, and the sizes are determined by a series of experiments using Korean–Chinese NMT. The trained model includes four hidden layers with 1000 nodes in each layer. Other hyperparameters followed the default OpenNMT settings (embedding dimensionality 512, beam size 5, dropout 0.7, batch size 32, and sgd optimizer with learning rate 1.0 for training). The validation perplexity converged to 5.54 at the end of 20 training epochs.

3.3. Feature-Based Classifier with an Automatically Expanded Dataset

Since creating human-labeled training sets for hybrid translation is time-consuming, expensive, and generally not available at large scales, we use support vector machines (SVMs) as the basic classifier, because SVMs are efficient, especially with a large number of feature dimensions and smaller datasets, while a small training set makes deep learning prohibitive. Since neural networks have shown good performance in classification using only word vectors, we also compare the feature-based classifier with several text classifiers to see whether the proposed features are still effective in such a comparison.

We use a self-training approach to construct the training corpus automatically. First, we trained the feature-based classifier 1 on the human-labeled dataset, and then we automatically constructed a larger dataset with the trained classifier. The automatically classified dataset is used with the human-labeled data to train the classifier 2 for hybrid translation, as shown in Figure 1 above.

4. Feature for Hybrid Machine Translation

The classifier must determine whether the NMT or RBMT translation is more reliable. Therefore, the features fed into the classifier are designed to reflect the translation qualities of both the RBMT system and the NMT model.

The RBMT translation procedure includes analysis, transfer, and generation; thus, analysis and transfer errors have a direct impact on translation quality.

4.1. High-Frequency Error Features

Ambiguity leads to high-frequency errors in analysis and transfer procedures. Analysis ambiguity includes morpheme ambiguity (segmentation ambiguity in Chinese), POS ambiguity, and syntactic ambiguity. Transfer ambiguity includes semantic ambiguity and word-order ambiguity. These features are language dependent, and they have often been described as grammar features in previous research. These features are black box features because they are extracted from the source sentences without morpheme or POS analysis. This paper proposes 24 features to reflect high-frequency errors related to ambiguities in Korean and Chinese. Some of these are as follows:

4.1.1. Morpheme and POS Ambiguities

In Korean, same surface morphological forms might have different root forms:

- Sal: sa (buy), or sal (live)
- Na-neun: nal-neun (flying), or na-neun ("I" as pronoun, or "sprout" as verb)

4.1.2. Syntactic Ambiguities

Case particles in Korean are often omitted, which causes case ambiguities among the subjective, vocative, and objective cases. The particles "-wa/-gwa", which mean 'with' or 'and' in English, cause syntactic ambiguities when they are used to connect two nouns. Below are two examples where the syntactic structures are determined by the word semantics (see (a) and (b) in Figure 2). In this case, there are analysis ambiguities, because word semantics are not considered during analysis.

- Nam-pyeon-**gwa** maek-ju jom haet-da (I had a beer **with** my husband).
- Chi-kin-**gwa** maek-ju jom haet-da (I had some chicken **and** beer).

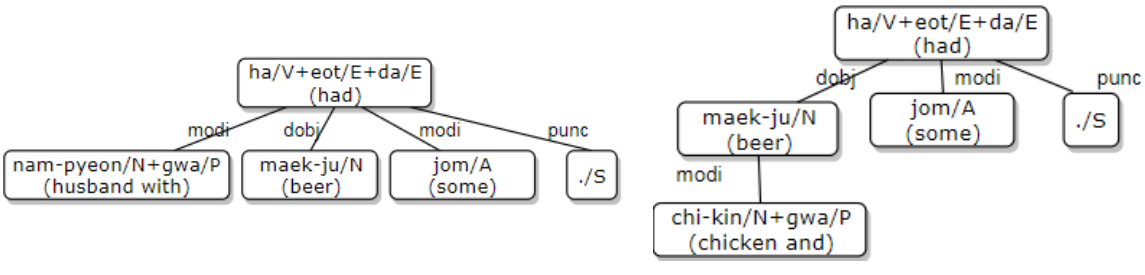

Figure 2. Syntactic ambiguity with "-gwa (with/and): (**a**) Dependency structure of the above first example; (**b**) Dependency structure of the above second example.

4.1.3. Ambiguity of Adverbial Postpositional Particles

Many adverbial postpositional particles in Korean not only have word-sense ambiguity, but they also cause word-order ambiguity in Korean-to-Chinese translation. This is seen with the high-frequency particles "-ro (with/by/as/to/on/onto)", "-e (at/on/to/in/for/by)", and "-kka-ji (even/until/by/to/up to)." The following are examples of "-ro" translation ambiguity:

- Taek-bae-**ro** mul-geon-eul bo-naet-da (I sent the goods **by** courier) → **Yong4** kuai4di4 fa1huo4
- Hoe-sa-**ro** mul-geon-eul bo-naet-da (I sent the goods **to** the company) → huo4wu4 ji4 **dao4** gong1si1 le

4.1.4. Word-Order Ambiguity Related to Long-Distance Reordering

There are some terms that are particularly relevant when discussing long-distance reordering in translation. In Korean, these include auxiliary predicates, such as "-ryeo-go_ha (want to)" and "-r-su_it (can/may)", and some verbs, such as "ki-dae_ha (expect)" and "won_ha (wish)". These words can cause word-order ambiguity during translation, which results in translation errors.

4.2. Basic Linguistic Features

Basic linguistic features have been widely used in previous research. These include POS, syntactic features, and surface features related to source sentences and target translations. These features are normally considered black-box features [22,23], but except for the length of the string feature, they can also be considered as glass-box features because they are produced by the POS tagger and parser included in the RBTM system [5].

4.2.1. Source- and Target-Sentence Surface Features

There are 19 features in this set. These include:

- Length features, such as the number of characters, morphemes, and syntax nodes. While the number of characters corresponds to the length of the string, the number of morphemes is equal to the tagged POS count. In some languages, the number of syntax nodes is the same as the number of morphemes, but they differ in Korean, because postpositional particles and ending words are not considered to be independent syntax nodes.
- Length comparison features of source and target sentences, such as the lengths of the NMT and RBMT outputs, the length ratio of each translation with the source sentence, and the length-difference ratio of each translation with the source sentence
- The number of unknown words in each translation and their proportion in the translation
- The number of negative words, such as "no/not", in the source sentence, and the difference in negative word counts between the source and each translation
- The number of numerals in the source sentence, and the difference in numeral counts between the source and each translation

4.2.2. Source-Sentence Linguistic Features

The number of different POSs (such as the number of verbs) and their proportion in the sentence. There are 46 total POS features.

The number of different syntax nodes (such as the number of objects or subjects) and their proportions in the syntactic nodes. There are 36 total syntactic features.

4.3. Pattern-Matching Features Reflecting Transfer Confidence

As previously introduced, the RBMT system uses transfer patterns to convert the source sentence into a translation, and the translation quality is strongly influenced by exact matches of the patterns. For example, if the original is translated with a long or informative pattern, it is more likely that the translation quality is higher than that of several short patterns or a less informative pattern.

There are 19 features proposed to reflect pattern-matching confidence, and all of them are glass-box features.

4.3.1. Pattern-Number Features

In the input sentence, adopted features include the number of words that need pattern matching, the number of the words matched with patterns, and the ratio of both. The lower the pattern-matching ratio, the lower the translation reliability. All words in Korean (except final endings and modal, tense-related words) must be matched with patterns for translation.

In the matched patterns, features include the number of arguments on the source side of the patterns and the proportion of matched arguments. When more arguments match with the input sentence, the match is more reliable.

4.3.2. Pattern-Weight Features

In the input sentence, adopted features include the total weight of the words that need pattern matching, the total weight of the words matched with patterns, and the ratio of these with the sentence.

In the matched patterns, adopted features include the weights of the patterns on the source side, and the ratio of the matched arguments' weights to the patterns' weights.

4.3.3. Pattern-Overlap and Pattern-Shortage Features

Pattern-overlap features include the number of verbs, auxiliary verbs, conjunctional connective words, and particles that match more than two patterns. If a word matches more than two patterns, the word position in those patterns may differ, and this can create word-order ambiguity related to word reordering, particularly for the above POSs. If these words do not match any patterns, this can also lead to translation errors.

Pattern-shortage features include the number of verbs, auxiliary verbs, connective words, and particles that fail to match any pattern.

4.3.4. High-Confidence Pattern Features

These features indicate whether an input sentence has been translated with a high-confidence pattern. There are patterns that are described at the sentence level, and most of their arguments are described with lexicon or sense tags. These patterns are considered high-confidence patterns. If an input sentence is translated using one of these patterns, the translation is usually trustworthy.

4.4. NMT Features

Most of the features in previous research that represent the translation quality of SMT, including the translation probability of target phrases and n-gram LM probability, were extracted from phrase tables and language models. Apart from the perplexity score of the final translation, NMT is more of a black box that barely produces sufficient information for feature extraction.

However, there are still factors that affect the quality of NMT translations, such as the number of numerals and rare words. We propose 20 NMT features as follows:

4.4.1. Translation-Perplexity Features

The perplexity produced by the NMT decoder for each translation can be adopted as a glass-box feature. The lower the perplexity, the more confident the model is regarding the translation.

The normalized perplexity score according to the translation is also adopted as a feature, because perplexity tends to be high when the target output is long.

4.4.2. Token-Frequency Features

Token-frequency classes are used for the source and target side. The average token frequency of each side is normalized to create 10 classes (1 to 10) that are adopted as black-box features. Tokens with higher frequencies belong to higher classes.

The average frequency ratio of the target tokens and the source tokens are captured using the ratio of their average frequencies and the ratio of their classes.

4.4.3. Token Features

To avoid the rare-word problem, NMT normally uses sub-words as tokens for both source and target sentences.

The numbers of numerals, foreign tokens, unknown tokens, and low-frequency tokens are captured with eight total black-box features for the source-sentence and its translation. Foreign tokens can be on either the source or target side (for example, English words in both a Korean source-sentence and a Chinese translation). Unknown tokens are those that are outside the token dictionary's scope or untranslated tokens on the target side (for example, untranslated Korean tokens remaining in the Chinese translation). Low-frequency tokens are tokens included in the token dictionary whose occurrence frequency in the training corpus is under a specified threshold.

The counts of mismatched numerals and foreign tokens in the source sentence and in the translation are captured with one feature. If both the source and target sentences have the same number of numerals or foreign tokens, but one of these numerals or foreign tokens differs in the translation (from that in the source sentence), we consider this a mismatch.

4.5. Rule-Based Features

By considering the proposed features above, several new rule-based features are obtained to detect whether the RBMT or NMT result is trustworthy. This produces four glass-box features in total.

4.5.1. RBMT Pattern Feature

Whether the RBMT result is unreliable is captured by considering pattern-matching features and calculating penalties. For example, if the ratio of matching pattern weights is below a given threshold, or if pattern matching is severely lacking, the RBMT result is not reliable.

4.5.2. Basic-Linguistic Features

Whether the RBMT result is unreliable is captured by considering the mismatching numerals and negative words along with other potential grammar errors.

Which translation is less reliable is captured by considering the number of unknown words in RBMT and NMT and whether the length difference ratio of RBMT or NMT exceeds a threshold. For example, if RBMT translates all words, but two words of NMT are not translated, the results of RBMT will be considered more trustworthy. Alternatively, if the original sentence and the RBMT translation are very long, but the NMT translation is very short, the NMT may not be trusted.

4.5.3. Classification Feature

After considering the rule-based features above, we add a feature to indicate whether we prefer the RBMT or NMT translation.

5. Experiments

We empirically evaluate the effects of the proposed hybrid translation in the following aspects:

- Translation performance and feature evaluation. The best feature set is selected through feature ranking and translation-accuracy estimation.

- Translation-performance evaluation with the auto-expanded corpus. The hybrid translation accuracy with an auto-expanded corpus is compared to the accuracy of RBMT and NMT.
- Classification performance comparison with text classification. Feature-based classification accuracy is compared with deep neural network text classification to determine whether the feature-based approach is still effective.

5.1. Experimental Settings

The first step is manually constructing the training data for the hybrid classifier. A set of source sentences are provided with the automatic translation results and the extracted features from RBMT and NMT. Human translators evaluate the quality of each translation result with a score (from 0 to 4 points) following the criteria shown in Table 2, which was revised from preliminary works [5,35]. Based on the human-estimated scores, each sentence is labeled with "NMT" (indicating that the NMT result is better), "RBMT" (indicating that the RBMT result is better), or "Equal". The extracted labeled features comprise the training corpus for classification.

Table 2. Scoring criteria for translation accuracy.

Score	Criterion
4	The meaning of the sentence is perfectly conveyed.
3.5	The meaning of the sentence is almost perfectly conveyed, except for some minor errors, such as incorrect articles or stylistic errors.
3	The meaning of the sentence is almost conveyed (approximately 75% of the sentence parts are translated correctly).
2.5	The meaning of the sentence is conveyed in part, but not in whole (approximately 60–70% of the sentence parts are correctly translated).
2	Some of the phrases in the sentence are translated correctly (approximately 50% of the sentence parts are correctly translated).
1	Only some words are correctly translated.
0	There is no translation or a completely incorrect translation.

The human-estimated scores are also adopted to estimate the translation engine's translation accuracy, as shown in Equation (1).

$$\text{translation accuracy} = \frac{\sum_{i=1}^{n}\left(\frac{\text{score}_i}{4}\right)}{n} \times 100.0 \quad (1)$$

Table 3 shows the classification corpus's composition by domain. The in-domain corpus contains 2786 sentences in the travel, shopping, and diary-conversation domains; these are the same domains as in the NMT training corpus, but they were developed independently. The out-of-domain corpus contains 1180 sentences collected from news, book reviews, Twitter, and lecture videos. The translation accuracy is estimated by domain.

Table 3. Information of hybrid classification corpus.

	Sentence	Average	Translation accuracy			The percentage of each label		
Domain	count	length	RBMT	NMT	Upper	RBMT	NMT	Equal
In domain	2,786	3.04	83.44%	92.84%	96.36%	8.33%	36.36%	55.31%
Out of domain	1,180	9.73	69.60%	67.44%	78.49%	38.73%	30.93%	30.34%
Total	3,966	5.05	79.32%	85.28%	91.00%	17.37%	34.75%	47.88%

As shown in Table 3, NMT shows better performance in in-domain translation, and RBMT shows better performance in out-of-domain translation; these results are consistent with the small-scale evaluation results shown in Table 1. We consider the NMT translation accuracy as our baseline translation accuracy, which is 85.28%. The baseline classification accuracy is 47.88%, which

is the percent of the sentences labeled "Equal". The upper bound of perfect classification would be 91.00%, as "Upper" column in Table 3.

We applied 10-fold cross-validation to the above datasets in the following experiments, unless otherwise noted.

5.2. Translation Performance and Feature Evaluation

All 181 features (f181) described in this paper are ranked with a feature selection tool provided for SVM [36]. Of these features, 152 are suggested as the best feature set (f152), and 76 are suggested as the second-best feature set (f76). We performed hybrid classification using structured SVMs [37] with the three feature sets above and compared the resulting translation accuracies to determine the final feature selection.

As shown in Table 4, the suggested best feature set of 152 features (f152) achieves the best translation accuracy (86.41%) in hybrid translation. This indicates that the feature-ranking result is reliable. Comparing these results with those in Table 3, we can see that regardless of the adopted feature set, the translation accuracy of the hybrid system exceeds that of the baseline NMT of 85.28%. For out-of-domain translation, the increase is even more obvious. The out-of-domain translation accuracies for RBMT and NMT are 69.60% and 67.44%, respectively, and the hybrid system's translation accuracy is 70.10%.

Table 4. Translation performance evaluation and feature selection.

Domain	f181	f152	f76
In domain	92.74%	93.31%	93.40%
Out of domain	69.59%	70.10%	69.72%
Total	85.86%	86.41%	86.36%

To determine whether the features proposed in this paper contribute to this performance, we have analyzed the top 30 features in the ranking, and we found that of the top 30 features, 43.33% originate in this study: seven features are RBMT pattern-matching features (Section 4.3), four features are NMT-related features (Section 4.4), and two features are rule-based features (Section 4.5) (Table 5). In the following experiments, the 152-feature set is adopted unless otherwise noted. Other "General" type features in Table 5 are from the high-frequency error features (Section 4.1) and basic linguistic features (Section 4.2).

Table 5. Feature ranking.

Rank	Feature type	Feature description
1	General	NMT perplexity score
2	General	RBMT translation word number
3	General	Source sentence string length (character number)
4	General	Source sentence word number
5	General	NMT translation word number
6	General	NMT translation length
7	Pattern matching	The word number needs to match the pattern
8	General	Source sentence morpheme number
9	General	Source sentence syntactic node number
10	Pattern matching	The word number matches the pattern
11	Pattern matching	The weight of words needs to match the patterns
12	NMT-related	The average frequency ratio of translation and source sentence
13	General	The syntactic node number modifier
14	General	The proportion of POS punctuations
15	NMT-related	The class of the average frequency of target tokens
16	NMT-related	Normalized NMT perplexity score

17	NMT-related	The class of the average frequency of source tokens
18	Pattern matching	The verb number needs to match the patterns
19	General	The verb number in a source sentence
20	Pattern matching	The weight of words in matched patterns
21	Pattern matching	The verb number in matched patterns
22	General	The length ratio of the target and source sentence
23	General	POS number – verb
24	General	POS number – ending
25	General	Syntactic proportion – punctuations
26	General	Syntactic proportion – modifier
27	General	POS number – particles
28	Rule-based	Rule – need NMT by penalty scores
29	Rule-based	Rule – need NMT by length ratio
30	Pattern matching	The number of overlapped patterns matching on particles

5.3. Translation Performance Evaluation with Auto-Expanded Corpus

Reference [5] proposed a truncation method to construct a training dataset for an SMT/RBMT combination using a classification approach. The truncation method operated as follows: if the SMT translation achieved a high enough confidence score (BLEU in their research), then the sentence was labeled "SMT"; otherwise, it was labeled "RBMT". This simple method was sufficient in their research. However, unlike the SMT/RBMT hybrid, in which the translation accuracy of each system is similar to the other, in the NMT/RBMT hybrid, the performance of NMT is much better than RBMT for in-domain translation and slightly lower for out-of-domain translation. We evaluated the truncation method on the same dataset used in Table 3, and we found that even the upper bound translation accuracy of the hybrid system with the best truncation could not outperform NMT.

We collected 20,000 sentences (which are independent from both the NMT training corpus and the human-labeled training corpus) for classification. We translated these sentences with both NMT and RBMT and performed feature extraction. We then labeled them automatically with the classifier trained on the human-labeled corpus. We used both the human-labeled corpus and the automatically labeled corpus to train a new classifier. This process is the same as that shown in Figure 1. The classifier with the automatically expanded dataset is compared with the classifier with the human-labeled dataset to see if the automatically expanded dataset contributes to the hybrid model's translation performance. We still apply 10-fold cross-validation to the dataset in Table 3, so in each iteration, we have 23,569–23,570 training examples and 397–396 test examples.

As shown in Table 6, the translation accuracy (T-accuracy) increased from 86.41% to 86.63% because of the improvement on the out-of-domain data, despite the in-domain translation accuracy, which declined slightly. Since, as seen in Table 3, 47.88% of the data are labeled "Equal" (indicating the NMT and RBMT translation qualities are the same), we also measure the classification accuracy (C-accuracy) to see how much the classification performance is influenced by the expanded dataset. As shown in the last row of Table 5, classification accuracy is greatly improved: from 54.46% to 63.01%, or +8.55%.

Table 6. Translation accuracy with auto-expanded training data. T-accuracy: translation accuracy, C-accuracy: classification accuracy.

Domain	Criteria	Human dataset	Expanded dataset
In domain	T-accuracy	93.31%	92.80%
Out of domain	T-accuracy	70.10%	71.75%
Total	T-accuracy	86.41%	86.63%
Total	C-accuracy	54.46%	63.01%

Compared to the increase in classification accuracy, the increase in translation accuracy is not very large. This is because the evaluation dataset is mostly composed of elements labeled "Equal"

and "NMT", as seen in Table 3. According to Equation (1), our classification formula, if the classification result is "Equal" or "NMT", the final translation choice is "NMT". Therefore, if the classifier incorrectly predicts the "NMT" case as "Equal", or vice versa, it has no effect on translation accuracy. Translation accuracy is only affected when "NMT" or "Equal" is incorrectly classified as "RBMT" or when "RBMT" is incorrectly classified as "NMT" or "Equal".

5.4. Comparison with Text Classification

For text-based models, we adopt a convolutional neural network text classifier (CNN-text) [38], a recurrent neural network (RNN) text classifier [39] with long short-term memory and gated recurrent units, and a BERT [40] multilingual model as the pre-trained model. The tokens from the source sentence and the translation outputs of RBMT and NMT are adopted as inputs for the text classifiers.

For our feature-based classifiers, the above RNN classifier (RNN), the above RNN classifier with the above CNN classifier (CNN+RNN), a multilayer perceptron model (MLP), and a CNN classifier (CNN) are adopted. The MLP model comprises a three-layer MLP with ReLU (256 units) and dropout (0.1) after each layer, followed by a softmax layer. The CNN feature-based classifier is composed of four layers: two convolutional 2D layers (filter size 100, kernel size 3 × 3, and ReLU activation function), a max-pooling layer (size 2 × 2), a feedforward layer with ReLU (256 units), and a softmax layer with dropout after max-pooling and feedforward layers. Both classifiers are optimized with the Adam algorithm. Early stopping is added to avoid overfit neural network-based models, and most models stopped in 10 epochs. Classification accuracy is used as the evaluation criterion for comparing classification models.

As shown in Table 7, the feature-based approach proposed in this paper clearly outperforms text-based classification, including with the BERT pre-trained model. This is because the features proposed in this paper express in-depth features explicitly, whereas text cannot express them well. In addition, feature-based classification obviously benefits from an expanded corpus, but text-based classification is the opposite. As we mentioned above, the expanded corpus is independent of the human-labeled training corpus, which is also used as the 10-fold cross-validation dataset. Again, it implies that the features proposed in this paper could explore the common in-depth features from two independent datasets.

Table 7. Comparing feature- and text-based classifiers. CNN: convolutional neural network, RNN: recurrent neural network, MLP: multilayer perceptron model, SVM: support vector machine.

Classifier	CNN-text	RNN	BERT	SVM	CNN	RNN	MLP	CNN+RNN
Inputs	Tokens	Tokens	Tokens	f.152	f.152	f.152	f.152	f.152
Human dataset	49.31%	52.82%	55.07%	54.46%	55.97%	51.94%	55.14%	54.54%
Expanded dataset	50.53%	50.27%	48.06%	63.01%	59.05%	53.70%	58.85%	58.78%

6. Conclusions

This paper deals with a classification-based hybridization that selects the best translation between the results of NMT and RBMT. Considering that previously used measures and features tend to evaluate frequency and fluency, which create a preference bias for NMT, we investigated RBMT-related features, including pattern-matching features and rule-based features that measure RBMT's quality, and we proposed NMT-related features that reflect NMT's quality. We also expanded the training data automatically using a classifier trained on a small, human-constructed dataset. The proposed classification-based hybrid translation system achieved a translation accuracy of 86.63%, which outperformed both RBMT and NMT. In practice, it is difficult to improve the accuracy of translation system after a certain degree (such as 80%). Thus, the above improvement is remarkable. The contribution to performance is more evident in out-of-domain translations, where RBMT and NMT showed translation accuracies of 69.60% and 67.44%, respectively, while the proposed hybrid system scored 71.75%.

Since deep-learning text classification and sentence classification have shown good performance, we also compared feature-based classification with text classification. Our experiments show that feature-based classification still has better performance than text classification.

Future studies will focus on low-resource NMT by exploring ways to utilize the in-depth and explicit linguistic knowledge that has been used in RBMTs system until now. We expect to use user dictionaries and syntactic information to improve NMT's ability to control low-frequency-word translation and sentence structure in long-sentence translation.

Author Contributions: Conceptualization, J.-X.H.; methodology, J.-X.H.; software, J.-X.H.; validation, J.-X.H.; formal analysis, J.-X.H.; investigation, J.-X.H.; resources, J.-X.H.; data curation, J.-X.H.; writing—original draft preparation, J.-X.H.; writing—review and editing, J.-X.H. and K.-S.L.; visualization, J.-X.H.; supervision, K.-S.L.; project administration, Y.-K.K.; funding acquisition, Y.-K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by "Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement)", "Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A1B03036275),", and "research funds of Jeonbuk National University in 2018".

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Du, J.; Way, A. Neural Pre-Translation for Hybrid Machine Translation. In Proceedings of the MT Summit XVI, Nagoya, Japan, 18–22 September 2017; pp. 27–40.
2. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the Association for Computational Linguistics, Philadelphia, PA, USA, 6-12 July 2002; pp. 311–318.
3. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, 25–27 June 2007; pp. 177–180.
4. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A.M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv* **2017**. arXiv:1701.02810.
5. Park, E.-J.; Kwon, O.-W.; Kim, K.; Kim, Y.-K. A Classification-based Approach for Hybridizing Statistical Machine Translation and Rule-based Machine Translation. *ETRI J.* **2015**, *37*, 541–550.
6. Dabbadie, M.; Hartley, A.; King, M.; Miller, K.J.; Hadi, W.M.E.; Popescu-Belis, A.; Reeder, F.; Vanni, M. A Hands-on Study of the Reliability and Coherence of Evaluation Metrics. In Proceedings of the LREC, Las Palmas, Canary Islands, Spain, 29–31 May 2002.
7. Dugast, L.; Senellart, J.; Koehn, P. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System, In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007.
8. Simard, M.; Ueffing, N.; Isabelle, P.; Kuhn, R. Rule-Based Translation with Statistical Phrase-Based Post-Editing. In Proceeding of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007.
9. Dugast, L.; Senellart, J.; Koehn, P. Statistical Post Editing and Dictionary Extraction: Systran/Edinburgh Submissions for ACL-WMT2009. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30–31 March 2009; pp. 110–114.
10. Schwenk, H.; Abdul-Rauf, S.; Barrault, L.; Senellart, J. SMT and SPE machine translation systems for WMT'09. In Proceedings of the Forth ACL Workshop on Statistical Machine Translation, Athens, Greece, 30–31 March 2009; pp. 130–134.

11. Lee, K.-Y.; Kim, Y.-G. Applying Statistical Post-Edit to English-Korean Rule-based Machine Translation System. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, Bali, Indonesia, 16–18 November 2012; p. 318.
12. Bojar, O.; Chatterjee, R.; Federmann, C.; Haddow, B.; Huck, M.; Hokamp, C.; Koehn, P.; Logacheva, V.; Monz, C.; Negri, M.; et al. Findings of the 2015 Workshop on Statistical Machine Translation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisboa, Portugal, 17–18 September 2015; pp. 1–46.
13. Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huck, M.; Yepes, A.J.; Koehn, P.; Logacheva, V.; Monz, C.; et al. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*; Association for Computational Linguistics: Berlin, Germany, 11–12 August 2016; pp. 131–198.
14. Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huang, S.; Huck, M.; Koehn, P.; Liu, Q.; Logacheva, V.; et al. Findings of the 2017 Conference on Machine Translation. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; pp. 169–214.
15. Bechara, H. Statistical Post-Editing and Quality Estimation for Machine Translation Systems. Master's Thesis, Dublin City University, Dublin, Ireland, 2014.
16. Chatterjee, R.; Weller, M.; Negri, M.; Turchi, M. Exploring the Planet of the APEs: A Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), Beijing, China, 26–31 July 2015; pp. 156–161.
17. Chatterjee, R.; Negri, M.; Rubino, R.; Turchi, M. Findings of the WMT 2018 Shared Task on Automatic Post-Editing. In Proceedings of the Third Conference on Machine Translation (WMT), Belgium, Brussels, 31 October–1 November 2018; pp. 710–725.
18. Avramidis, E.; Popovic, M.; Burchardt, A. DFKI's Experimental Hybrid MT System for WMT 2015. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; pp. 66–73.
19. Ueffing, N.; Ney, H. Application of Word-level Confidence Measures in Interactive Statistical Machine Translation. In Proceedings of the Ninth Annual Conference of the European Association for Machine Translation (EAMT-2005), Budapest, Hungary, 24–25 November 2005; pp. 262–270.
20. Huang, F.; Papineni, K. Hierarchical System Combination for Machine Translation. In Proceedings of the Empirical Methods in Natural Language Processing, Prague, Czech Republic, 28–30 June 2007; pp. 277–286.
21. Specia, L.; Cancedda, N.; Dymetman, M.; Turchi, M.; Cristianini, N. Estimating the Sentence-Level Quality of Machine Translation Systems. In Proceedings of the 13th Annual Conference of the EAMT, Barcelona, Spain, 14–15 May 2009; pp. 28–35.
22. Specia, L.; Shah, K.; Souza, J.G.C.; Cohn, T. QUEST—A Translation Quality Estimation Framework. In Proceedings of the 51st ACL, Sofia, Bulgaria, 4–9 August 2013; pp. 79–84.
23. Specia, L.; Paetzold, G.H.; Scarton, C. Multi-Level Translation Quality Prediction with QUEST++. In Proceedings of the ACL-IJCNLP 2015 System Demonstrations, Beijing, China, 26–31 July 2015; pp. 115–120.
24. Federmann, C. Hybrid Machine Translation Using Joint, Binarised Feature Vectors. In Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas, San Diego, CA, USA, 28 October–1 November 2012; pp. 113–118.
25. Avramidis, E. Comparative Quality Estimation: Automatic Sentence-Level Ranking of Multiple Machine Translation Outputs. In Proceedings of the COLING, Mumbai, India, 8–15 December 2012.
26. Shah, K.; Cohn, T.; Specia, L. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In Proceedings of the XIV MT Summit, Nice, France, September 2–6 2013; pp. 167–174.
27. Hunsicker, S.; Yu, C.; Federmann, C. Machine Learning for Hybrid Machine Translation. In Proceedings of the 7th Workshop on Statistical Machine Translation, Montréal, QC, Canada, 7–8 June 2012; pp. 312–316.
28. Li, H.; Zhao, K.; Hu, R.; Zhu, Y.; Jin, Y. A Hybrid System for Chinese-English Patent Machine Translation. In Proceedings of 6th Workshop on Patent and Scientific Literature Translation (PSLT6), Urumqi, China, 25–26 August 2015; pp. 52–67.
29. Eisele, A.; Federmann, C.; Uszkoreit, H.; Saint-Amand, H.; Kay, M.; Jellinghaus, M.; Hunsicker, S.; Herrmann, T.; Chen, Y. Hybrid Machine Translation Architectures Within and Beyond the Euromatrix

Project. In Proceedings of the 12th Annual Conference of the European Association for Machine Translation, Hamburg, Germany, 22–23 September 2008; pp. 27–34.
30. Federmann, C.; Eisele, A.; Chen, Y.; Hunsicker, S.; Xu, J.; Uszkoreit, H. Further Experiments with Shallow Hybrid MT Systems. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics, Uppsala, Sweden, 11–16 July 2010; pp. 77–81.
31. Espana-Bonet, C.; Labaka, G.; Ilarraza, A.D.; Marquez, L.; Sarasola, K. Hybrid Machine Translation Guided by a Rule-Based System. In Proceedings of the 13th Machine Translation Summit, Xiamen, China, 19–23 September 2011; pp. 554–561.
32. Dhariya, O.; Malviya, S.; Tiwary, U.S. A Hybrid Approach for Hindi-English Machine Translation. In Proceedings of the 2017 International Conference on Information Networking (ICOIN), Da Nang, Vietnam, 11–13 January 2017.
33. Yin, C.H.; Seo, Y.A.; Kim, Y.-G. Korean-Chinese Machine Translation Using Three-Stage Verb Pattern Matching. In Computer Processing of Oriental Languages. Language Technology for the Knowledge-Based Economy, ICCPOL 2009, Hong Kong, China, 26–27 March 2009; Li, W., Mollá-Aliod, D., Eds.; Lecture Notes in Computer Science; Berlin: Springer, 2009; Volume 5459.
34. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th ACL, Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
35. Kwon, O.-W.; Choi, S.-K.; Lee, K.-Y.; Roh, Y.-H.; Kim, Y.-G. Customizing an English-Korean Machine Translation System for Patent/Technical Documents Translation. In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, China, 3–5 December 2009; pp. 718–725.
36. Chen, Y.W.; Lin, C.J. Combining SVMs with Various Feature Selection Strategies. In *Feature Extraction, Foundations and Applications*. Berlin: Springer, 2006; pp.315–324.
37. Lee, C.; Jang, M.-G. A Modified Fixed-Threshold SMO for 1-Slack Structural SVMs. *ETRI J.* **2010**, *32*, 120–128.
38. CNN Text Classifier. Available online: https://github.com/likejazz/cnn-text-classification-tf (accessed on 6 December 2019).
39. RNN Text Classifier. Available online: https://github.com/jiegzhan/multi-class-text-classification-cnn-rnn (accessed on 6 December 2019).
40. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.