Very Large-Scale Lexical Resources to Enhance Chinese and Japanese Machine Translation

Jack Halpern
The CJK Dictionary Institute, Inc.

Abstract

A major issue in machine translation (MT) and natural language processing (NLP) applications is the recognition and translation of named entities. This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. This paper focuses on some of the linguistic issues related to orthographic variation in Japanese, the challenges of translating Japanese named entities, and conversion to Traditional Chinese. It also introduces several Very Large-Scale Lexical Resources (VLSLR) designed to significantly enhance translation accuracy, and argues that the quality of neural machine translation (NMT) systems can be significantly enhanced through the integration of lexicons.

1. Introduction

1.1 Major issues

A major issue in MT and other NLP applications is the recognition and translation of named entities. This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. These difficulties are exacerbated by the lack of easily-available comprehensive lexical resources, especially for named entities, and the lack of a standardized orthography in Japanese. Some of the major factors that contribute to the difficulties of Chinese and Japanese NLP in general, and MT in particular, include:

1. Since the Japanese orthography is highly irregular, identifying, disambiguating and normalizing the large number of orthographic variants requires support for advanced capabilities such as cross-script normalization (Halpern, 2008).
2. The morphological complexity of Japanese requires the use of a robust morphological analyzer, rather than a simple *n*-gram tokenizer, to perform such operations as segmentation, lemmatization, and decompounding (Brill et al., 2001; Yu et al., 2000).
3. The accurate conversion between Simplified Chinese (SC) and Traditional Chinese (TC), a deceptively simple but in fact extremely difficult computational task (Halpern and Kerman, 1999).
4. The difficulty of accurately translating POIs points of interest, such as schools, highways, hotels, etc.).
5. The difficulty of performing accurate segmentation of Japanese and Chinese texts (Goto et al., 2001; Yu et al., 2000), which requires identifying word boundaries by breaking a text stream into meaningful semantic units.
6. Miscellaneous technologies such as the recognition of lexeme (rather than morpheme) and discontinuous multiword MWEs (e.g. extracting 'take off' + 'jacket' from 'he took his

jacket off"), synonym expansion, and cross-language information retrieval (CLIR) (Goto et al., 2001).

7. The lack of easily available comprehensive lexical resources for named entities such as proper nouns, especially POIs. Such entities pose special difficulties, as they are extremely numerous, difficult to detect without a lexicon, and have an unstable orthography.

1.2 Goals of this paper

Each of these issues deserves a paper in its own right. Here we will focus on some key linguistic issues related to Japanese and Chinese MT, including (1) the typology of orthographic variation in Japanese, (2) the challenges of translating Japanese POIs, (3) the difficulty of converting to Traditional Chinese accurately, and (4) how the integration of lexicons into MT systems, especially NMT systems, can overcome these difficulties.

The paper also introduces several Very Large-Scale Lexical Resources (VLSLR) consisting of millions of CJK named entities, such as a multilingual database of Japanese POIs and personal names, a very comprehensive multilingual database of Chinese personal names, and large-scale bilingual technical term databases for Chinese and Japanese, and argues that such resources can significantly enhance translation accuracy for both traditional MT systems and for state-of-the-art NMT systems.

2. Japanese Orthographic Variants

2.1 Irregular Orthography

One reason that the Japanese script is difficult to process by NLP tools and MT systems is its highly irregular orthography. The numerous orthographic variants result from, among other factors, the unpredictable interaction between the four scripts used in Japanese; namely, kanji, hiragana, katakana and the Latin script. This can be illustrated by the sentence 金の卵を産む鶏 /Kin no tamago wo umu niwatori/ 'A hen that lays golden eggs.' . Tamago 'egg' has four variants (卵，玉子，たまご，タマゴ), /niwatori/ 'chicken' has three (鶏，にわとり，ニワトリ) and /umu/ 'give birth to' has two (産む，生む), which expands to 24 permutations. Since these variants occur frequently, MT and NLP systems have no hope of identifying them as instances of the same underlying sentence without support for orthographic disambiguation/normalization.

2.2 Variant Typology

There are eight types orthographic variation in Japanese (Halpern, 2008). The three most important ones are described below.

1. ***Okurigana variants.*** This refers to kana endings attached to a kanji base or stem, such as *okonau* 'perform', written 行う or 行なう, whereas *atarihazure* can be written in the six ways shown in the table below. Identifying and normalizing *okurigana* variants, which are numerous and unpredictable, is a major issue. An effective solution is to use an orthographic variants lexicon, a solution adopted by some major search engine portals.

Atarihazure	Type of variant
当たり外れ	"standard" form
当り外れ	okurigana variant
当外れ	okurigana variant
当外	all kanji
当たりはずれ	replace kanji with hiragana
あたり外れ	replace kanji with hiragana

Table 1. Variants of *atarihazure*

2. ***Cross-script variants.*** This refers to variation across the four Japanese scripts in Japanese, including hybrid words written in multiple scripts, as shown below. Cross-script variants, which are common and unpredictable, negatively impact recall and pose a major challenge to NLP applications, including MT.

Kanji	Hiragana	Katakana	Latin	Hybrid	English
人参	にんじん	ニンジン			carrot
		オープン	OPEN		open
硫黄		イオウ			sulfur
		ワイシャツ		Y シャツ	shirt
皮膚		ヒフ		皮フ	skin

Table 2. Cross-script variants

3. ***Katakana variants.*** The use of katakana loanwords is a major annoyance in MT since katakana words are very numerous and their orthography is often irregular. It is common for the same word to be written in multiple, unpredictable ways, as shown below:

Type	English	Standard	Variants
Macron	computer	コンピュータ	コンピューター
Long vowels	maid	メード	メイド
Multiple kana	team	チーム	ティーム

Table 3. Katakana variants

3. Chinese Orthographic Variants

Below is a brief description of the major issues in Chinese orthographic variation.

3.1 Simplified vs. Traditional Chinese

In mainland China and Singapore, the characters are written in simplified forms called Simplified Chinese (SC), whereas Taiwan, Hong Kong, Macau and most overseas Chinese communities continue to use the old, complex forms referred to as Traditional Chinese (TC). Several factors contribute to the complexity of the Chinese script: (1) the large number of

characters, (2) the major differences between SC and TC along various dimensions (graphemic, semantic and phonemic), (3) the many orthographic variants in TC, and (4) the difficulty of accurately converting between SC and TC.

3.2 Chinese-to-Chinese Conversion

The process of automatically converting SC to/from TC, referred to as "C2C," is fraught with pitfalls. A detailed description of the linguistic issues can be found in (Halpern and Kerman, 1999), while the technical issues related to encoding and character sets are described in Lunde (2008). The conversion can be implemented on three levels in increasing order of sophistication.

Code conversion. The most unreliable way to perform C2C conversion is on a codepoint-to-codepoint basis by looking up in a mapping table, such as the one below. Because of the numerous one-to-many mappings (which occur in both the SC-to-TC and the TC-to-SC directions), the rate of conversion failure is unacceptably high.

SC	TC1	TC2	TC3	TC4
语	語			
松	鬆	松		
干	幹	乾	干	榦

Table 4. Code conversion

Orthographic conversion. The next level of sophistication in C2C is to convert orthographic units: that is, meaningful linguistic units, especially compounds and phrases that match on SC to TC characters on a one-to-one basis. This gives better results because the orthographic mapping tables enable conversion on the word or phrase level rather than the codepoint (character) level.

English	SC	TC	Comment
country	国家	國家	correct
change	变化	變化	correct
relax	放松	放鬆	correct
Caracas	加拉加斯	加拉加斯	should be 卡拉卡斯
Yemen	也门	也門	should be 葉門

Table 5. Orthographic conversion

The ambiguities inherent in code conversion can be resolved by using orthographic mapping table like the above, but because there are no word boundaries ambiguities must be resolved with the aid of a segmenter that can break the text stream into meaningful units (Emerson, 2000).

Lexemic conversion. A more sophisticated, and more challenging, approach to C2C conversion is to map SC to TC lexemes that are semantically, rather than orthographically, equivalent. For example, SC 信息 (*xìnxī*) 'information' is converted to the semantically

equivalent TC 資訊 (*zīxùn*). This is similar to the difference between *lorry* in British English and *truck* in American English.

There are numerous lexemic differences between SC and TC, especially in technical terms and proper nouns (Tsou et al., 2000). To complicate matters, the correct TC is sometimes locale-dependent, as shown below. Lexemic conversion is the most difficult aspect of C2C conversion, and can only be done with the help of mapping tables.

English	SC	Taiwan TC	Other TC
software	软件	軟體	軟件，軟件
taxi	出租汽车	計程車	的士，德士
Osama Bin Laden	奥萨马本拉登	奧薩瑪賓拉登	奧薩瑪賓拉丹
Yemen	也门	葉門	
Caracas	加拉加斯	卡拉卡斯	

Table 6. Lexemic conversion

4. Lexicons in MT

4.1 Lexicons in traditional MT

Lexicons, including dictionary databases and terminology glossaries, have played a critical role in NLP applications in general, and in MT systems in particular. There is no question that large-scale lexicons have dramatically improved translation quality in traditional MT systems, especially in view of the fact that these systems perform rather poorly on out-of-domain texts (Mediani et al., 2014).

Attempts to replace lexicons with algorithmic solutions for certain tasks, such as processing Japanese orthographic variants and katakana loanwords, have been made (Brill et al., 2001). To successfully process the highly irregular orthography of Japanese, *lexeme*-based procedures such as orthographic disambiguation cannot be based on probabilistic methods alone. Many attempts have been made along these lines, as for example in Brill et al. (2001) and Goto et al. (2001), with some claiming performance equivalent to lexicon-based methods, while Kwok (1997) reports good results with only a small lexicon and simple segmentor.

In fact, such algorithmic/statistical methods have only met with limited success. The fundamental problem is that such methods, even when based on large-scale corpora, often fail to achieve the high accuracy required for NLP and MT applications unless they are supported by large-scale lexicons. For example, Emerson (2000) and Nakagawa (2004) and others have shown that MT systems and robust morphological analysers capable of processing lexemes, rather than bigrams or *n*-grams, must be supported by a large-scale computational lexicons (even 100,000 entries is much too small).

4.2 Quantum leap

The application of artificial neural network to MT gave birth to a new paradigm, Neural Machine Translation (NMT), that can be said to represent a quantum leap in translation technology. In a short period of time, such major MT engines as Google, Bing and Baidu adopted the NMT model, whose success can be attributed to its capability to implement the

translation process on the basis of a single, end-to-end probabilistic model (Luong et al., 2015).

Even as NMT development proceeds at breakneck speed, greatly contributing to translation quality, research on newer advanced technologies based on Quantum Neural Networks (QNN) is already in progress (Moire et al., 2016). However, as we shall see below, despite of the significant improvement in translation quality, the of ability of NMT systems to correctly translate named entities and some technical terms has somewhat deteriorated.

4.3 Lexicons in NMT

On April 25-26, 2017 the TAUS Executive Forum Tokyo 2017 (TAUS, 2017) was held in Tokyo, where the team leaders and representatives of several major NMT developers (Google, Microsoft, NICT) gathered. Several papers were presented on the current state of NMT technology and speech-to-speech translation. In discussions with several NMT experts, including Chris Wendt from Microsoft and Manuel Herranz from Pangeanic, it became clear that though currently the major NMT systems do not use lexicons, there is no technical reason that lexicons cannot be integrated into NMT systems.

The basic idea is to regard a lexicon as a kind of sentence-aligned, bilingual parallel corpus, and to have the system assign a higher probability to the lexicon entries so as to override the results of the normal NMT algorithms. For example, 三角線 /Misumi-sen/, the name of a railway line in Kyushu, is called 'Misumi Line' in English, so that it is safe to allow the lexicon results to override the NMT results such as 'Triangle' (Google) and 'Triangular line' (Bing).

Some potential obstacles are (1) that lexicons, unlike corpora, do not provide context, and (2) that ordinary lexicons do not provide translation probabilities. However this is not critical for named entities, especially POIs, and even for many technical terms, since named entities are mostly monosemic (have only one word sense), which means that word sense disambiguation is unnecessary and that the lexicon can automatically be assigned a higher probability. For example, there is no danger that 三角線 should be correctly translated literally as 'triangular line'. rather than 'Misumi Line', the official name of this train line.

4.4 Lexicon integration

NMT has transformed MT technology by achieving significant quality improvements over traditional MT systems. When NMT systems are trained on large-scale domain-specific parallel corpora, they do achieve remarkable results *within* those domains.

According to Arthur et al. (2016), NMT does not perform well when "translating low-frequency content words that are essential to understanding the meaning of the sentence." Our experiments (see §5 below) have confirmed that NMT systems also perform poorly when translating named entities, especially POIs, as well as when processing Japanese orthographic variants. Arthur et al. (2016) propose that this can be overcome by integrating "discrete translation lexicons" into NMT systems, and assert that the accuracy of probability can be improved by leveraging information from discrete probabilistic lexicons. They go on to discuss the difference between "automatically learned lexicons" and "manual lexicons," and how these can be integrated into NMT systems, and conclude that as a result of incorporating discrete probabilistic lexicons into NMT systems "we achieved substantial increases in BLEU (2.0-2.3) and NIST (0.13-0.44) scores, and observed qualitative improvements in the translations of content words."

In summary, although the major NMT systems do not currently incorporate lexicons, it is clear that with some effort they can be configured to do so. It is also clear that integrating lexicons into NMT systems is highly desirable since it will lead to major improvements in translation quality. Ideally, NMT should take advantage of the positive aspects of SMT, and even RBMT, and merge them into new kind of hybrid system that offers the best of both worlds.

5. Experiments and Results

Both traditional MT systems as well as state-of-the-art NMT systems often fail to recognize and accurately process named entities such as Japanese proper nouns, especially POIs. Below are the results of some spot tests we conducted using three major NMT engines, namely Google Translate, Bing Translate, Baidu Translate, and NICT's TextTra (a phrase-based SMT system), on Japanese POIs, Japanese orthographic variants, Traditional Chinese conversion and Chinese technical terms, and comparing the results with our own (CJKI's) large-scale proper noun databases.

5.1 Japanese Points of Interest

Our tests to translate 75 Japanese POIs (with focus on railway lines, airports and amusement facilities) into English using the two major US NMT engines gave surprisingly poor results.

Japanese	Google	Bing	CJKI
海の中道線	Midair line of the sea	The middle line of the sea	Umi-no-Nakamichi Line
三角線	Triangle	Triangular line	Misumi Line
十和田観光電鉄線	Towada Shimbun photoelectric wire	Towada Kanko railway line	Towada Kanko Electric Railway Line
神津島空港	Kozu Island airport	God Tsushima Airport	Kozushima Airport
中部国際空港	Chubu International Airport	Chubu International Airport	Chubu Centrair International Airport
鬼の城公園	Demon Castle Park	Demon Castle Park	Oninojo Park

Table 7. POIs by Google and Bing

Using the major Asian engines (Baidu and NICT) for the same POIs gave the following results:

Japanese	Baidu	NICT	CJKI
海の中道線	The sea line	海の中道線	Umi-no-Nakamichi Line
三角線	Misumi	Misumi Line	Misumi Line
十和田観光	Towada sightseeing	Towada Kankō	Towada Kanko

電鉄線	electric railway line	Electric Railway Line	Electric Railway Line
神津島空港	Kozu Island Airport	Kōzushima Airport	Kozushima Airport
中部国際空港	Central Japan International Airport	Chubu International Airport	Chubu Centrair International Airport
鬼の城公園	Demon Castle Park	Oni Castle Park	Oninojo Park

Table 8. POIs by Baidu and NICT

5.2 Evaluation of results

Our institute (CJKI) uses five methods to determine the level of accuracy of POI translation. In principle, these five levels represent increasing accuracy and increasing production costs.

1. *Transliteration* (字訳) refers to representing the source script (graphemes, not phonemes) with the characters of another script, as in AR محمد → \mHmd\ or JN 幕張国際展示場 → ZH 幕张国际展示场.
2. *Phonemic transcription* (音訳) which represents the phonemes of the source language, as in AR محمد→ romanized *muHammad* and JN 東京中央ゴルフ場 → romanized *Tokyo Chuo Gorufujo*.
3. *Semantic-phonemic transcription* (意音訳) combines semantic transcription with phonemic transcription, as in JN 東京中央ゴルフ場 → EN *Tokyo Chuo Golf Course.*
4. *Semantic transcription* (意訳) refers to semantically converting (translating) components into the target language, as in JN 幕張国際展示場 → 幕张国际展览馆 and JN 東京中央ゴルフ場 → *Tokyo Central Golf Course* (e.g., JN 展示場 is equivalent to ZH 展览馆).
5. *Human translation* (翻訳) is converting to the correct semantic equivalent (the "official" name) in the target language, such as JN 幕張国際展示場 → ZH 幕张国际展览中心 and JN 東京中央ゴルフ場 → EN *The Central Golf Club, Tokyo.*

The first four can be done algorithmically by referencing component mapping tables and a conversion rules database; that is, semiautomatically with some human proofreading. The fifth, the highest level, can be done accurately only by looking up in hand-crafted lexicons, such as CJKI's proper noun databases, which for years have served as the gold standard in the Named Entities Workshop (NEWS) transliteration task (Zhang, et al., 2012).

The success rate for the above MT engines is less than 50%. "Success" is defined as level 5 above, meaning that the results should be (almost) identical to the entries in CJKI's POI databases (ignoring minor differences such as long vowels), which have been manually proofread to ensure accuracy. The results are summarized below:

Google	47%
Microsoft	40%
Baidu	39%
NICT	47%

Comparing these results to CJKI's, it is clear that some errors result from translating the POI components literally (semantic transcription), rather than the named entity as a whole. For example, 鬼の城公園 was translated as 'Demon Castle Park' since the string 鬼の城 consists of 鬼の 'demon' + 城 'castle', whereas the actual name of this park in English is 'Oninojo Park'. That is, 鬼の城公園 was not recognized as a named entity but was translated literally component by component.

5.3 Orthographic variation

It seems as if NMT engines do not perform orthographic normalization or disambiguation for Japanese, and probably not for other languages as well. Since Japanese has a highly irregular and unstable orthography, this has a major negative impact on Japanese translation quality. Let's consider the orthographic variants for the following three words:

English	Reading	Var. 1	Var 2	Var. 3
sun	hi	日	陽	
mansion	yashiki	屋敷	邸	
shine	sasu	差す	さす	射す

Table 9. Typical variants in Japanese

This means that a sentence like /hi no sasanai yashiki/ 'a mansion that gets no sunshine' can have such variants as:

日の差さない屋敷	日の差さない屋敷	陽の射さない屋敷
日の射さない屋敷	日の差さない邸	陽の差さない屋敷
日のささない屋敷	日のささない邸	陽のささない屋敷
陽の射さない邸	陽の差さない邸	陽のささない邸

Table 10. Highly irregular Japanese orthography

Running some of these through Google and Bing we get:

Japanese	Google	Bing
日の差さない屋敷	A dwindling residence	A house with no sun
日の射さない屋敷	A mansion that does not shine.	She mansion of the day.
日のささない屋敷	A daydreaming residence.	A mansion with no sun
陽のささない屋	A ya man who does not sunlight.	A house with no sunshine

Table 11. Japanese variants by Google and Bing

An analysis shows (1) that though these phrases are 100% equivalent, they are being considered as distinct, and (2) that no orthographic normalization takes place. For example, Google translated 陽 /hi/ 'sun' to the mysterious 'ya man' and is not aware that it is an orthographic variant and 日 /hi/ 'sun'. In the case of Bing, 'A house with no sunshine' is 100%

correct, but 'She mansion of the day' makes no sense at all. Baidu and NICT give similarly poor results:

Japanese	Baidu	NICT
日の差さない屋敷	There's no day at home	Residence that deprive Japan of.
日の射さない屋敷	Day without sunshine house.	Residence not days.
日のささない屋敷	Deprive of the residence.	Residence which do not refer to date.
陽のささない屋敷	The residence where no.	The mansion where the sun never bites.

Table 12. Japanese variants by Baidu and NICT

It is interesting to note that NICT often interprets 日 as 'date' or 'day', rather than the correct 'sun'. Here too there are some translations that make no sense, such as Baidu's 'There's no day at home' and NICT's 'Residence that deprive Japan of'. Clearly, none of the MT engines surveyed is doing orthographic normalization, which is critical for Japanese.

5.4 Traditional Chinese

Below are the results of translating into TC by the four MT engines:

ENG	SC	CJKI	Google	Bing	Baidu	NICT
computer	计算机	電腦	計算機	電腦	電腦	計算機
database	数据库	資料庫	數據庫	資料庫	資料庫	資料庫
file	文件	檔案	文件	檔	檔案	檔案
information	信息	資訊	信息	資訊	資訊	信息
software	软件	軟體	軟件	軟體	軟件	軟體
Taxi	出租车	計程車	出租車	計程車	計程車	齣租車

Table 13. Comparing lexemic conversion

Comparing the results to CJKI (the gold standard based on CJKI's lexemic conversion tables), the results are almost 100% right for Baidu (a Chinese company) and Bing. Google, on the other hand, is clearly strictly limited to orthographic conversion, so that 信息, for example, is incorrectly translated to the orthographic 信息 rather than the lexemic 資訊. For 'taxi' NICT gives the mysterious 齣租車 instead of the intended orthographic 出租車, which correctly should be 計程車. Clearly, NMT systems could benefit from using lexemic mapping tables.

5.5 Technical terminology

Translation quality of MT systems depends on such factors as the size and quality of the training corpus, the MT model and algorithms, and supporting lexicons. Some systems, such as NICT's, have been trained on patent corpora and thus achieve good accuracy in patent translation (Sumita, 2013). Our spot checks have confirmed that NMT engines do perform better in the domains of science and technology than in translating named entities such as POIs.

Nevertheless, the lack of technical terminology lexicons does have a negative impact on NMT systems. For example, comparing CJKI's large-scale Chinese technical term databases (five million entries) demonstrates that the NMT results are mostly incorrect results for some medical terms, as shown below:

Chinese	CJKI	Google	Bing	Baidu	NICT
类天花	alastrim	smallpox	class smallpox	smallpox like	smallpox
类骨质	osteoid	bone-like	bone type	osteoid	bone
孢子丝菌病	sporotrichosis	spore mycosis	spore silk fungus disease	histoplasmosis	Spore 丝菌病
亚氨基酸	imino-acid	amino acid	amino acids	imino acid	亚氨基酸
亚硫酐	sulfurous anhydride	aulfurous acid	arian	sulfurous anhydride	亚硫酐

Table 14. Technical terms by the four NMT engines

6. Lexical Resources

6.1 Very Large-Scale Lexical Resources

The CJK Dictionary Institute (CJKI), which specializes in CJK and Arabic computational lexicography, has for decades been engaged in research and development to compile comprehensive lexical resources, with special emphasis on dictionary databases for CJK and Arabic named entities, technical terminology, and Japanese orthographic variants, referred to as Very Large-Scale Lexical Resources (VLSLR). Below are the principal resources designed enhance the accuracy of MT and NLP applications.

6.2 Japanese resources

1. The *Japanese Personal Names Database* covers over five million entries, including hiragana readings, numerous romanized variants (sometimes over 100 per name) and their English, SC, TC, and Korean equivalents.
2. The *Japanese Lexical/Orthographic Database* covers about 400,000 entries, including *okurigana,* kanji, and kana variants for orthographic disambiguation and grammar codes for morphological analysis.
3. The *Comprehensive Database of Japanese POIs and Place Names,* which covers about 3.1 million entries in 14 languages along with hiragana and romanized variants.
4. The *Database of Katakana Loanwords* covers about 50,000 entries.
5. The *Database of Japanese Companies and Organizations* covers about 600,000 entries.

6.3 Chinese resources

1. The *Comprehensive Chinese to Chinese Mapping Tables* (C2C) exceeds 2.5 million entries. This covers general words, named entities and technical terms mapped to their TC equivalents, including such attributes as POS codes and type codes, and supports all three conversion levels, namely code, orthographic and lexemic conversion.
2. The *Database of 100 Million Chinese Personal Names,* an extremely comprehensive resource (under construction), covers Chinese personal names, their romanized variants,

dialectical variants for Cantonese, Hokkien and Hakka, multilingual coverage for English, Japanese, Korean, and Vietnamese.

3. The *Database of Chinese Full Names* covers 4 million Chinese full names of real people, including celebrities.
4. Miscellaneous mapping tables such TC orthographic normalization tables and large-scale pinyin databases showing the difference between SC and TC pronunciation, and others.

7. Conclusions

With computer memory being inexpensive and virtually unlimited, it is no longer necessary for traditional MT systems to over-rely on corpora and algorithmic solutions. The time has come to leverage the full power of large-scale lexicons to significantly enhance the accuracy of NLP applications in general, and MT systems in particular. As for NMT, although the major engines do not currently incorporate lexicons, it is clear that the effort to do so is highly desirable since it will lead to major improvements in translation quality. Although "lexicon integration" does pose some technical challenges, it is a worthwhile goal to pursue and deserves the serious attention of NMT researchers and developers. Ideally, a new kind of "hybrid NMT" that leverages the power of traditional MT systems combined with neural networks should be developed.

References

Arthur, P., Neubig, G. and Nakamura, S. (2016). Incorporating Discrete Translation Lexicons into Neural Machine Translation. In *Proceedings of EMNLP 2016: Conference on Empirical Methods in Natural Language Processing*, Austin, Texas.

Brill, E., Kacmarcik, G. and Brockett, C. (2001). Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 393-399, Tokyo, Japan.

Emerson, T. (2000). Segmenting Chinese in Unicode. In *Proceedings of the 16th International Unicode Conference*, Amsterdam

Goto, I., Uratani, N. and Ehara, T. (2001). Cross-Language Information Retrieval of Proper Nouns using Context Information. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 571-578, Tokyo, Japan

Halpern, J. and Kerman, J. (1999). The Pitfalls and Complexities of Chinese to Chinese Conversion. In *Proceedings of the Fourteenth International Unicode Conference*, Cambridge, MA.

Halpern, J. (2008). Exploiting Lexical Resources for Disambiguating Orthographic CJK and Arabic Orthographic Variants. In *Proceedings of LREC 2008*. Marrakesh, Morocco.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, Cambridge, MA.

Kwok, K.L. (1997). Lexicon Effects on Chinese Information Retrieval. In *Proceedings of the 2nd Conference on Empirical Methods in NLP*. ACL, 141-148, Stroudsburg, PA.

Luong, M., Pham, H. and Manning, C.D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, Lisbon, Portugal.

Lunde, K. (2008). *CJKV Information Processing*. O'Reilly & Associates, Sebastopol, CA.

Mediani, M., Winebarger, J. and Waibel, A. (2014). Improving In-Domain Data Selection For Small In-Domain Sets. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA.

More, S., Dhir, G.S., Daiwadney, D. and Dhir, R.S. (2016). Review on Language Translator Using Quantum Neural Network (QNN). *International Journal of Engineering and Techniques*, Volume 2 Issue 1, Jan - Feb 2016, Chennai, India.

Nakagawa, T. (2004). Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. In *Proceedings of the 20th international conference on Computational Linguistics*, p.466-es, Geneva, Switzerland.

Sumita, E. (2013). Multi-Lingual Translation Technology: Special-Purpose System for Multi-Lingual High-Quality Translation. *Journal of the National Institute of Information and Communications Technology*, pages 35-39, Tokyo, Japan.

TAUS Executive Forum Tokyo. (2017). https://www.taus.net/events/conferences/taus-executive-forum-tokyo-2017. Retrieved May 8, 2017.

Tsou, B.K., Tsoi, W.F., Lai, T.B.Y., Hu, J. and Chan, S.W.K. (2000) LIVAC, A Chinese Synchronous Corpus, and Some Applications. In *Proceedings of the ICCLC International Conference on Chinese Language Computing*, pages 233-238, Chicago.

Yu, S., Zhu, X. and Wang, H. (2000). New Progress of the Grammatical Knowledge-base of Contemporary Chinese. *Journal of Chinese Information Processing, Institute of Computational Linguistics*, Peking University, Vol.15 No.1.

Zhang, M., Li, H., Liu, M. and Kumaran, A. (2012). Whitepaper of NEWS 2012 shared task on machine transliteration. In *Proceedings of the 4th Named Entity Workshop (NEWS '12)*. Association for Computational Linguistics, Stroudsburg, PA.