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

Jack Halpern
The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
jack@cjki.org

Abstract

A major issue in machine translation (MT) 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 discusses some of the major issues in Japanese and Chinese MT, such as the difficulties of translating proper nouns and technical terms, and the complexities of orthographic variation in Japanese. Of special interest are neural machine translation (NMT) systems, which suffer from a serious out-of-vocabulary problem. However, the current architecture of these systems makes it technically challenging for them to alleviate this problem by supporting lexicons. This paper introduces some Very Large-Scale Lexical Resources (VLSLR) consisting of millions of named entities, and argues that the quality of MT in general, and NMT systems in particular, can be significantly enhanced through the integration of lexicons.

Keywords: machine translation, lexicon, Chinese, Japanese

1. Introduction

A major issue in MT applications is the translation of named entities and technical terms. This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. Some factors that contribute to these difficulties include:

1. The Japanese orthography is highly irregular, requiring advanced capabilities such as cross-script normalization (Halpern, 2008).
2. The morphological complexity of Japanese requires the use of a robust morphological analyzer for segmentation and lemmatization (Brill et al., 2001; Yu et al., 2000).
3. The accurate conversion between Simplified Chinese (SC) and Traditional Chinese (TC) (Halpern and Kerman, 1999).
4. The difficulty of accurately translating POIs (points of interest). These are extremely numerous, difficult to detect, and have an unstable orthography.
5. The large number of technical terms.
6. The lack of comprehensive lexical resources.

This paper discusses some of these issues, and introduces Very Large-Scale Lexical Resources (VLSLR) consisting of millions of named entities. It argues that lexicons can enhance the translation accuracy of NMT systems, which currently don't use lexicons.

2. Japanese Orthographic Variants

The Japanese orthography is highly irregular. The numerous orthographic variants, which are common and unpredictable, negatively impact recall and pose a major challenge to MT. The variation results from the unpredictable interaction between four scripts: kanji, hiragana, katakana, and Latin. For example, in 金の卵を産む鶏 '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. Algorithmic solutions have no hope of identifying these as instances of the same underlying sentence without support for orthographic disambiguation/normalization.

The most important types of orthographic variation in Japanese (Halpern, 2008) are: (1) *Okurigana*, which are kana endings such as 当たり外れ and 当り外れ for *atarihazure*. Normalizing *okurigana* variants, which are numerous and unpredictable, is a major issue. An effective solution is to use an orthographic variants lexicon. (2) *Cross-script variants* refers to variation across the four Japanese scripts, 'carrot' (*ninjin*) written in kanji (人参), hiragana (にんじん), and katakana (ニンジン). (3) *Katakana loanword variants*, a major annoyance since they are numerous and irregular. The same word to be written in multiple, unpredictable ways, such as コンピュータ and コンピューターa for 'computer' and チーム and ティーム for 'team'.

3. Lexicons in MT

3.1 Lexicons in traditional MT

Lexicons, including dictionary databases and terminology glossaries, have played a critical role in MT systems, dramatically improving translation quality, 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 Japanese orthography of Japanese orthographic disambiguation cannot be based on probabilistic methods alone. Attempts have been made along these lines, as for example in Brill 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 high accuracy MT 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).

3.2 Quantum leap

The application of artificial neural network to MT gave birth to a new paradigm, Neural Machine Translation (NMT), that represents a quantum leap in MT 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, research on newer advanced technologies based on Quantum Neural Networks (QNN) is already in progress (Moire et al., 2016). However, despite of the significant improvement in translation quality, the ability of NMT systems to correctly translate named entities and some technical terms has in fact somewhat deteriorated.

3.3 Lexicons in NMT

According to He et al. (2016) of Baidu, "an NMT system usually has to apply a vocabulary of certain size... thus it causes a serious out-of-vocabulary problem." Baidu is probably the only company that has tackled the difficult problem of integrating lexicons into MT systems.

On April 25-26, 2017 the TAUS Executive Forum Tokyo 2017 (TAUS, 2017) was held in Tokyo, and on September 18-22, 2017 MT Summit XVI was held in Nagoya, Japan, where the team leaders and representatives of several major NMT developers (Google, Microsoft, NICT) gathered. In discussions with several NMT experts, including Chris Wendt from Microsoft and representatives of Baidu, it became clear that though currently the major NMT systems (except for Baidu) do not use lexicons, there is no technical reason that lexicons cannot be used. 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, is called 'Misumi Line', 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, 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.

3.4 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 §4 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 asserts 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 (except for Baidu), do not currently incorporate lexicons, 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 merge them into new kind of hybrid system that offers the best of both worlds.

4. Experiments and Results

Both traditional MT systems as well as state-of-the-art NMT systems often fail to accurately translate Japanese proper nouns, especially POIs. Below are the results of some spot tests using three major NMT engines, namely Google Translate, Bing Translate, Baidu Translate, and NICT's TextTra (phrase-based), on Japanese POIs, Japanese orthographic variants, and Chinese technical terms, and comparing the results with CJKI's large-scale terminology databases.

4.1 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
鬼の城公園	Demon Castle Park	Demon Castle Park	Oninojo Park

Table 1. 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
鬼の城公園	Demon Castle Park	Oni Castle Park	Oninojo Park

Table 2. POIs by Baidu and NICT

4.2 Evaluation of results

Our institute (CJKI) uses five methods to determine the level of accuracy of POI translation, in increasing order of accuracy. (1)*Transliteration* refers to representing the source script in another script, as in JN 幕張国際展示場 to ZH 幕张国际展示场, (2) *phonemic transcription*, representing the phonemes of the source language, as in romanizing JN 東京中央ゴルフ場 to *Tokyo Chuo Gorufujo*, (3) *semantic-phonemic transcription* combines semantic transcription with phonemic transcription, as in JN 東京中央ゴルフ場 translated to EN *Tokyo Chuo Golf Course*, (4) *semantic transcription* translates components into the target language, as in JN 幕張国際展示場 to ZH 幕张国际展览馆 and JN 東京中央ゴルフ場 into EN *Tokyo Central Golf Course*, and (5) *human translation*, which is translating to the correct semantic equivalent (the "official" name), such as JN 幕張国際展示場 to ZH 幕张国际展览中心 and JN 東京中央ゴルフ場 to 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 have served as the gold standard in the Named Entities Workshop (NEWS) transliteration task (Zhang, et al., 2012).

The success rate for the four MT engines tested was less than 50% (Google 47%, Microsoft 40%, Baidu 39%, and NICT 47%). "Success" is defined as level 5 above, meaning that the results should be (almost) identical to the entries in CJKI's POI databases, which have been manually proofread. 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 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.

4.3 Orthographic variation

It seems as if NMT engines do not perform orthographic normalization or disambiguation for Japanese. 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 3. 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, 日の差さない邸, 陽の差さない屋敷 and 陽の射さない邸.

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 4. 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 of 日 *hi* 'sun'. For Bing, 'A house with no sunshine' is 100% correct, but 'She mansion of the day' makes no sense. 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 5. Japanese variants by Baidu and NICT

Note that NICT often interprets 日 as 'date' or 'day', rather than '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.

4.4 Technical terminology

Translation quality depends on such factors as the size and quality of the training corpus, the MT model and algorithms, and supporting lexicons. Despite the dramatic contributions of NMT to translation quality, the problem of unknown vocabulary remains (He et al., 2016), especially for names entitied like POIs and the huge number of technical terms. Some systems, such as NICT's, have been trained on patent corpora and thus achieve good accuracy in patent translation (Sumita, 2013).

Chinese	CJKI	Google	Bing	Baidu	NICT
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
亚硫酐	sulfurous anhydride	aulfurous acid	arian	sulfurous anhydride	亚硫酐

Table 6. Technical terms by four NMT engines

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. For example, comparing CJKI's Chinese technical term databases (millions of entries) demonstrates that the NMT results are often incorrect for some medical terms, as shown in Table 6.

5. Lexical Resources

5.1 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 to enhance MT quality.

5.2 Japanese resources

1. The *Japanese Personal Names Database* covers over five million entries, including hiragana readings, numerous romanized variants 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.
4. The *Database of Katakana Loanwords.*

5.3 Chinese resources

1. The *Comprehensive Simplified Chinese to Traditional 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.
4. Miscellaneous mapping tables such as large-scale pinyin databases showing the difference between SC and TC pronunciation, and others.

6. 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. As for NMT, although most engines do not currently incorporate lexicons, clearly the effort to do so is desirable since it will lead to major improvements in translation quality. Although "lexicon integration" does pose technical challenges, it is a worthwhile goal 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.

7. Bibliographical 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

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.

He, W., He, Z, Wu, H., and Wang, H. (2016). Improved Neural Machine Translation with SMT Features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, AZ.

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.