Adapting Chinese Word Segmentation for Machine Translation Based on Short Units

Yiou Wang, Kiyotaka Uchimoto, Jun'ichi Kazama, Canasai Kruengkrai, and Kentaro Torisawa

National Institute of Information and Communications Technology,

3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

E-mail:{ wangyiou,uchimoto,kazama,canasai,torisawa }@nict.go.jp

Abstract

In Chinese texts, words composed of single or multiple characters are not separated by spaces, unlike most western languages. Therefore Chinese word segmentation is considered an important first step in machine translation (MT) and its performance impacts MT results. Many factors affect Chinese word segmentations, including the segmentation standards and segmentation strategies. The performance of a corpus-based word segmentation model depends heavily on the quality and the segmentation standard of the training corpora. However, we observed that existing manually annotated Chinese corpora tend to have low segmentation granularity and provide poor morphological information due to the present segmentation standards. In this paper, we introduce a short-unit standard of Chinese word segmentation, which is particularly suitable for machine translation, and propose a semi-automatic method of transforming the existing corpora into the ones that can satisfy our standards. We evaluate the usefulness of our approach on the basis of translation tasks from the technology newswire domain and the scientific paper domain, and demonstrate that it significantly improves the performance of Chinese-Japanese machine translation (over 1.0 BLEU increase).

1. Introduction

An important feature of Chinese texts is that they do not explicitly indicate word boundaries by spaces. Word segmentation is therefore considered an important first step in MT, and its performance has a great impact on MT results. In fact, a substantial amount of research has been carried out to provide solutions to Chinese word segmentation problems in MT. However, most research has focused on either using the correspondence between Chinese words and words that are explicitly tokenized in the target language (e.g., English) (Cao et al., 2009; Ma & Way, 2009; Xu et al., 2004) or combining various segmenters in statistical machine translation (SMT) training or decoding (Dyer et al., 2008; Zhang et al., 2008). One important and yet often neglected factor is the optimal segmentation granularity of the manually segmented monolingual corpus, which is widely used by corpus-based segmenters. One exception is an investigative study showing that the granularity of Chinese words is very important in MT (Chang et al., 2008). Beyond this, no detailed analysis exists concerning what style and size of word unit is optimal for MT. Moreover, most discussion related to the unit sizes of word segmentation for other NLP tasks, such as word alignment and information retrieval, has only considered whether to join or split noun compounds (Peng et al., 2002; Bai et al., 2008)

In this paper, we improve the machine translation quality by adjusting the granularity of the words and present a simple yet effective approach to adapt Chinese word segmentation to SMT based on short units. The basic idea is very simple: we define the short-unit segmentation standard and then transform the annotated corpus, which is used as the training data for the segmentation model, into the short-unit standard one by using a semi-automatic method, which requires human knowledge to add rules and modify the database at some extent. To evaluate the efficacy of our method, we conducted experiments on Chinese-Japanese SMT tasks from two corpora: technology newswire corpus and scientific paper corpus. The experimental results indicate that our approach improves the SMT performance (over 1.0 BLEU increase) in both corpora.

2. Short Units for Chinese Word Segmentation

2.1 Segmentation Unit

Word segmentation refers to the process of dividing a sentence into meaningful units called "word units" and word units greatly depend on the manually annotated corpus for a corpus-based word segmentation model. However we found the following word segmentation problems in existing manually annotated Chinese corpora, such as the famous CTB corpus (Penn Chinese Treebank) and PKU corpus (PKUTreebank, see Section 3.1):

(1) The segmentation unit of the existing corpora is not very clear and tokens are in different granularity levels: morpheme level, word level, and multi-word level.

For example, prefix 总 (*chief*) is treated as one segmentation, while compound words 旧民主主义革命 (*old democratic revolution*) are also treated as one segmentation.

(2) The segmentation standards tend to emphasize functional independency rather than phonological or morphological independency.

For example, consider the sentence, "丹尼尔先生/设计

/管理信息系统 (Mr. Daniel/ *design /management information system.).”* Here, the corresponding Chinese of *management information systems* is a compound noun and acts as the sentence's object. It is treated as one segmentation in the existing corpus, although it includes three morphologically independent words: 管理(management)/信息(information)/系统(system).

(3) The segmentation standards tend to have low segmentation granularity and provide poor morphological information.

For instance, Chinese multi-character words composed of more than one meaningful morpheme may be translated into several English words. For example, the Chinese word 艺术节 translates into *art festival* in English. Morphemes 艺术 and 节 have their own meanings and translate into *art* and *festival* respectively. However they are considered as one unit in both the CTB and the PKU corpus segmentation standards. Although 艺术节 is a known word in the training corpus, 艺术团 (a*rt troupe*), which shares the common morphological information 艺术 with known word 艺术节, was found out to be an out-of-vocabulary (OOV) word and could not be correctly analyzed upon experiments. However if 艺术节 is segmented into two words, 艺术/节, at least the knowledge of 艺术 can be learned from the training data and correctly segmented. Segmentations at a low granularity level are bound to increase out-of-vocabulary (OOV) words, which is the primary factor causing mis-segmentation. Lexicalization mismatch due to mis-segmentation degrades the performance of word alignment, whose quality greatly impacts on SMT performance.

Based on the above findings, we define the segmentation unit with high granularity -a shorter word unit relative to the existing standard and hereafter referred to as short unit- to be the smallest string of characters , with a relatively independent meaning, such as the “艺术”, “团”, and “节” in the previous example.

We assume that short units have the following merits:

(1) Because the shorter a word unit is, the more likely the token is the basic word, short units will alleviate the data sparseness problem.

(2) The short-unit standard is more stable than the existing standard: unifying the word granularity into one level will bring the high precision to the corpus.

(3) Short units are the effective unit for the example collection and statistical analysis (Ogura et al., 2009).

2.2 Segmentation Principles for Short Units

We propose the following principles to provide a procedural algorithm for identifying short units or to transform the segmentation in an existing corpus into short units.

(1) A two-character string in the existing annotated corpus should be treated as a short unit.

A Chinese character may have very different meanings in different contexts or in different collocations. Hence two-character words in Chinese are usually non-compositional and should not be decomposed into character units based on the definition of the short unit. Moreover word-unit level segmentation enhances the ability to disambiguate in MT. Therefore we retain two-character segmentations in the existing corpus.

For example, two-character Chinese word 上班 means “be on duty” or “go to work”. However if it is split into character units, 上 has such POS tags as verb, adjective, localizer and may mean “up”, “go to”, “climb”, “above” , “previous” and so on; 班 may mean “class”, “a (work) shift”, “a scheduled run of a public transportation” , “a squad” and so on. Splitting it into character units may cause ambiguity.

Some cases have no ambiguity, such as 南北 (south and north). Although 南 (south) and 北 (north) have independent meanings without ambiguity, we presently keep them in their original segmentation form. We will address such cases in future work.

(2) A string (three characters or more), whose meaning can be derived from the sum of its components, should be treated as the sum of the short units and split into short units, provided that there is no overlap between the short units and at least one of the short units is a multi-character string. For example, for a three-character string (ABC), if the meaning of the string=AB+C or =A+BC, then decompose it into short units AB/C or A/BC; if the meaning of the string=A+B+C or =AC+BC, then keep the string in its original segmented form.

For example, 上海市 (=上海+市) should be split into 上海/市, while 大中小 (=大+中+小) and 中小学 (=中学+小学) should be kept .

(3) A short string (less than three characters) with a high frequency should be treated as a short unit when necessary. We will explain it in detail in Section 2.3

(4) Strings separated by overt segmentation markers (such as punctuation marks) should be segmented, such as the result of a sporting even (e.g., 3:2 => 3/ :/ 2).

2.3 Short-Unit Transformation Approach

We accomplished the segmentation standard transformation (from the existing standard into the short-unit standard) through the following strategies:

(1) Designing transfer rules by referring to the Japanese short units and using the alignment results.

Our basic idea is to decompose the long-unit words (hereafter referred to as long tokens) into short units using the information acquired by performing Chinese-Japanese word alignment. We previously built a Chinese-Japanese bilingual corpus using sentences from the annotated PKU corpus as the source. At the same time, there is a Japanese short-unit dictionary called UniDic[(1)], which can be used by the MeCab[(2)] ,a Japanese morphological analyzer. We first segmented each Japanese sentence in the parallel corpus using MeCab along with UniDic into Japanese short-unit word sequences, and then used the GIZA++ toolkits to obtain the Chinese-Japanese word alignment

results. We extracted the 1-to-n alignment results to acquire the segmentation information of the long tokens and tried to automatically decompose such words. However we found that only a minority of long tokens, which should be split into short units, can be suitably decomposed in this manner. The reasons are as follows:

① The alignment result is not clean, containing a lot of noise and many errors.

② Some long tokens even with correct alignment information should not be decomposed based on the corresponding Japanese.

③ A long token that should be split based on our definition may correspond to one Japanese word, so there is no 1-to-n alignment information for the token.

The objective of short units is to make corpora with finer granularity and more consistent from a monolingual point of view. Therefore we generated transfer rules to split the long tokens into short units. We still used the 1-to-n Chinese-Japanese alignment pairs. The long token cases are categorized by part of speech. For each category, we designed short units and constructed transfer rules from human knowledge. Finally, we implemented the transfer rules by a program to transform long segmentations into short units. At present, there are 28 transfer rules.

For example, we define short units for number, such as "万" , "1-digit number + 千","1-digit number + 百", design the transfer rule for numbers and then implement the rule by a program. Long token " 三千八百二十四万二千一百二十一"can be transformed into short units "三千/八百/二十/四/万/二千/一百/二十/一".

(2) Building a transfer database by utilizing external lexicons

Some words have complex internal structures, such as nouns or verbs. Such long tokens cannot be transformed into short units by transfer rules. However since we know that all essential linguistic knowledge is encoded in the lexicon, we can utilize an external dictionary to obtain knowledge for short-unit transformation and build a short-unit transfer database. More specifically, we build the transfer database by adopting the following procedure:

Step 1 Compile the lemma list and the morpheme list given as follows:

(i) The lemma list (157,293 lemma entries) includes the following:

① proper nouns

② 2~4-character entries extracted from the external lexicon

We merged several electrical Chinese dictionaries and tokens in the existing corpora into one external lexicon that includes the word and POS-tag information. We extracted lemmas and morphemes from this external lexicon.

③ 2~3-character short lemmas derived from the lexicon

We exploited the short lemma and morpheme in the following method.

Based on the segmentation principle (3) (Section 2.2), we assume that a productive unlexicalized item (not an entry in the lexicon) has a relatively independent meaning and can be treated as a short unit. An unlexicalized item is productive, only if it can form several known words (entries in the lexicon), when combined with other known words. Based on this assumption, we segmented the entries with 3-4 characters in the lexicons into "known word + unlexicalized item" sequences and extracted the unlexicalized items with the frequency above three. (We here only deal with 3-4 character entries, because the entries with five or more characters are almost always compound nouns and proper nouns.)

For example, entries 一团和气 (very harmonious), 一团乱麻 (very chaotic), and 一团漆黑 (very dark) are segmented into "一团+和气", "一团+乱麻", and "一团+漆黑". Here 漆黑 (dark), 乱麻 (chaos), and 和气 (harmony) are known words, thus unlexicalized item "一团" is extracted as a short unit.

With the above method, we can exploit unlexicalized short units. The 2-3 character short units, such as 一团 in the previous example, are extracted as lemma entries, and the one-character short units are extracted as morpheme entries.

(ii) The bound morpheme list (549 morpheme entries) includes the following:

① prefixes (e.g.超,副,非,半…);

② suffixes (e.g.化,性,儿…);

③ bound morphemes for verbs, adjectives and adverbs (e.g., propositions 于, auxiliary morphemes 着, negative morphemes: 未,没...);

④ bound morphemes for proper nouns (e.g.人,山,市,队…)

⑤ bound morphemes for general nouns (e.g.节,机,色,….)

Step 2 Design the morpheme sequence pattern based on segmentation principle (2) (Section 2.2)

Pattern: (morpheme)*(lemma) +(morpheme)*

Here, the lemma is an entry in the lemma list (i), and the morpheme is an entry in the bound morpheme list (ii).

Step 3 Extract the segmentations in the annotated corpus that match the above pattern and generate a transformation database.

Step 4 Validate and correct the transfer database with human judgment to guarantee the high precision of the corpus and to solve the problem of segmentation ambiguity.

Step 5 Decompose long segmentations into short units based on the manually modified transfer database. At present there are 3617 unique long tokens in the transfer database. Some examples are shown in Table 1. The POS tag information is generated from the external dictionary and is manually modified in Step 4.

By using transfer rules and the transfer database, the manually annotated corpus is transformed into a short-unit standard one. Figure 1 shows the flowchart of the

short-unit transformation approach for the PKU corpus. 18,585 long tokens in the PKU corpus that contains 405,147 tokens (4.6%) have been transformed into short units, among which 6343 tokens are decomposed by transfer rules and the rest by transfer database. Moreover about 86% tokens with three characters or more in the PKU corpus are decomposed by our method, and the remaining ones are usually indecomposable. Because the whole procedure is based on a manually segmented corpus, the manual check is not time-consuming. It may take one person 3~4 days to check the transfer database for a 30,000-sentences corpus.

Long token	Tag	Short-unit 1	Tag	Short-unit 2	Tag
意大利语 (Italian)	n	意大利 (Italy)	ns	语 (language)	n
全球化 (globalization)	v	全球 (global)	n	化 (~lization)	sv
全世界 (whole world)	n	全 (whole)	a	世界 (world)	n
决定于 (depend on)	v	决定 (depend)	v	于 (on)	p
标志着 (is representing)	v	标志 (represent)	v	着 (Be ~ing)	u
仅次于 (only next to)	v	仅 (only)	d	次于 (next to)	v
专家系统 (expert system)	n	专家 (expert)	n	系统 (system)	n

Table 1: Examples of transfer database

3. Experiments and Discussions

3.1 Experimental Setting

(1) Chinese word segmentation models

The Chinese word segmentation model used in our experiments is a discriminative word-character hybrid model for joint Chinese word segmentation and POS tagging (Kruengkrai et al, 2009), namely, MMA.

(2) Chinese monolingual corpus data

The above model is trained on the Treebank created by Peking University called PKUTreebank, Which contains 30,686 sentences.

(3) SMT model

The MT system used in this paper is Moses, a state-of-the-art phrase-based system (Koehn et al., 2003). We tuned the parameters with MERT (Och, 2003) on the devset (shown in the Table 2).

(4) Bilingual corpus data

The bilingual data we used in our experiments are from two domains: technology newswires (NIKKEI_BP: Nikkei BP science and technology newswire corpus) and scientific papers (NICT_JC_SP: NICT Japanese-Chinese scientific paper corpus). We divided the corpora into the train-set, devset, development test-set and blind test-set. We treated two test sets equally in the experiments, so hereafter referred to them as test-set1 and 2. The corpora's various statistics of are shown in Table 2.

3.2 Results and Evaluation

In the experiments, we compared the following three segmentation models:

(1) Character-unit model: Each Chinese character is interpreted as a single word.
(2) Baseline model: We trained the Chinese word segmentation model MMA with the original PKU corpus and adopted it as the baseline model
(3) Short-unit model: We trained the Chinese word segmentation model MMA with the PKU corpus in the short-unit standard and adopted it as the short-unit model.

Then we used the character-unit model, the baseline model, and the short-unit model as the Chinese segmenters in Chinese-Japanese SMT tasks and conducted Chinese-Japanese SMT experiments. The translation performance of the three models is shown in Table 3.

BLEU	NICT_JC_SP		NIKKEI_BP	
	Testset 1	Testset 2	Testset 1	Testset 2
character-unit	26.81	27.09	32.12	32.77
baseline	27.01	27.55	31.55	32.97
short-unit	28.29	28.70	33.35	34.10

Table 3: Performance of Chinese-Japanese SMT

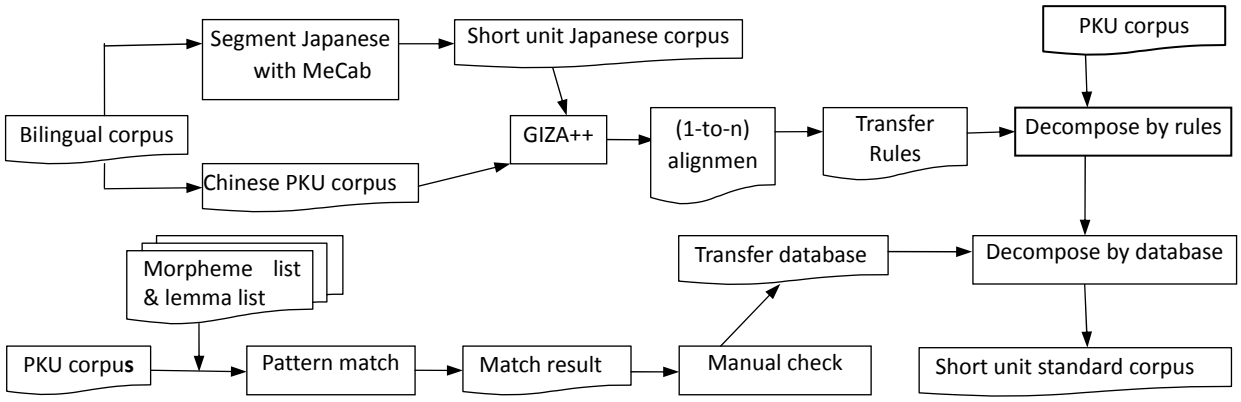

Figure 1: Flowchart of short-unit transformation approach

		Train-set		Devset		Testset1		Testset2	
		Chinese	Japanese	Chinese	Japanese	Chinese	Japanese	Chinese	Japanese
NIKKEI_BP	# sentences	245,554		1,000		500		500	
	#words	7,019,359	8,238,960	28,019	33,025	14,477	16.893	13,886	16,042
NICT_JC_SP	#sentences	371,868		1,603		500		500	
	# words	11,054,040	13,617,437	49,855	60,440	15,230	18.298	15,464	18,534

Table 2: Statistics of domain corpora

The character-unit MT in fact performs reasonably well and there is not a large gap between the character-unit and the baseline performance. However, the baseline, which can also be called the baseline word-unit, still outperforms the character-unit in all MT tasks, except for one test-set, and the short-unit method achieves significant improvement (over 1.0 BLEU increase) in SMT performance over both the baseline and the character-unit methods in both corpora. These experimental results suggest that the short-unit method achieves an optimal trade-off between character-unit and baseline word-unit segmentation for Chinese-Japanese SMT.

3.3 Discussions

(1) Effect of Segmentation on Translation Results

We present one example to show the effect that segmentation has on translation quality. Table 4 shows a segmented Chinese source sentence using different segmentation models, the corresponding SMT results, and the human reference translation.

In the example, the short-unit model provides correct segmentation and also gives a better translation, but segmentation errors in the baseline model lead to a completely wrong translation. Missegmentation is the main reason for the translation failure. For the character-unit segmentation, Moses extracts the “phrases" from word alignment, and the system constructs the useful words, so most words are correctly translated. However, it still causes the translation ambiguity problem. The last two characters, which should be translated as “accept”, are translated as “receive a message”.

(2) Why is the short-unit method effective for SMT?

We believe that the main point lies in the following aspects:

(i) Provides more consistent segmentation for MT

We transformed the existing corpus into a short-unit one that makes the monolingual corpus more consistent by unifying it into one granularity level: the short-unit level. For a corpus-based segmenter, more consistent training data will lead to more consistent segmentation. Consistent segmentation is helpful for MT performance (Chang et al. 2008).

(ii) Affects the follow-up phrase extraction by changing the word alignment

(iii) Decreases the number of the potential OOVs.

We analyzed the OOV rates of both word segmentation and machine translation experiments and found that, compared with the baseline, the short-unit method decreased the number of the potential OOVs in both experiments.

In Table 5, we list some statistics of both short-unit and baseline segmenters in machine translation experiments. While segmenting the MT data, the short-unit model generates a smaller MT lexicon but provides a lower OOV rate than the baseline segmenter. A further analysis showed that most OOVs (over 60%) in the short-unit MT tests are also OOVs in the baseline MT tests and many of the rest OOVs in the baseline MT tests are words consist of three or four characters or even more. Usually they should be segmented into shorter words.

Short-unit segmented input: 尽管 整体 性能 不 太 高 ， 也 可 接受 。 (Although /as a whole/ performance/ not/ very/ high ,/also /can / accept)
Translation with short-unit model: それほど高くないにもかかわらず, 全体としての性能を受けることも可能である 。 (although it is not so high, it is possible to take the whole performance)
Baseline segmented input: 尽管 整体性 能 不 太 高 ， 也 可 接受 。 (Although/ entirety / can/ not/ very/ high/, /also/ can/ accept)
Translation with baseline model: にもかかわらず， 全体でもそれほど高くない 。 (Although, the entirety is not so high.)
Character-unit segmented input: 尽 管 整 体 性 能 不 太 高 ， 也 可 接 受 。
Translation with character-unit model: にもかかわらず, 全体の性能が高いことも受信できる 。 (Although, it can receive a message that the whole performance is high.)
Reference: 全体としてのパフォーマンスが高くなくても受け入れてしまいます。 (Even though the whole performance is not high, it can be accepted.)

Table 4: Example for three translation models

NICT_JC_SP corpus	#MT training lexicon size	#MT testset 1 lexicon size	MT testset 1 lexicon OOV rate	#MT testset 2 lexicon size	MT testset 2 lexicon OOV rate
baseline	175,395	1894	0.0797	1910	0.0838
short-unit	125,115	1728	0.0601	1730	0.0589

NIKKEI_BP corpus	#MT training lexicon size	#MT testset 1 lexicon size	MT testset 1 lexicon OOV rate	#MT testset 2 lexicon size	MT testset 2 lexicon OOV rate
baseline	168,558	3445	0.0662	3281	0.0634
short-unit	115,441	3064	0.0509	2914	0.0491

Table 5: Machine translation Lexicon statistics and potential OOV

NICT_JC_SP	Testset 1				Testset 2			
Training data size	1/10	1/5	1/2	1	1/10	1/5	1/2	1
baseline	24.38	24.35	27.10	27.01	26.37	27.00	27.67	27.55
short-unit	25.18	26.40	27.70	28.29	26.36	27.37	28.05	28.70
diff	0.80	2.05	0.60	1.28	-0.01	0.37	0.38	1.25

NIKKEI_BP	Testset 1				Testset 2			
Training data size	1/10	1/5	1/2	1	1/10	1/5	1/2	1
baseline	25.27	27.01	28.13	31.55	25.96	27.48	30.04	32.97
short-unit	26.58	28.06	30.46	33.35	27.23	29.06	31.79	34.10
diff	1.31	1.05	2.33	1.80	1.27	1.58	1.75	1.13

Table 6: Comparison of SMT performances with different sizes of training data

(3) Will more training data weaken the short-unit method?

To find the relation between corpus size and the effect of the short-unit method, we conducted the experiments with different sizes of training data. The SMT performance in BLEU is shown in Table 6. From the experimental results, we cannot conclude that if we have more training data, the short-unit method will have less effect on MT performance. Even for the one test-set, with the increase of the training data size, the improvement of MT performance achieved by the short-unit method increases correspondingly.

4. Conclusions and Future Work

We introduced simple strategies based on the short unit concept to make a Chinese word segmentation model effectively learn the morphological information from the training corpus and thereby improve the SMT performance. We experimentally evaluated our approach on translation tasks, showed that our simple method gives quite good results for Chinese-Japanese MT, and confirmed our hypothesis that short units provide more appropriate segmentation granularity for SMT. Moreover, our method's simplicity makes it suitable for processing large corpora at low cost.

For future work, we plan to construct a more systematic and comprehensive short-unit standard, apply the short-unit transformation method to corpora with different segmentation standards and combine the corpora into short-unit standard one to produce a much larger dataset for better training. We also plan to test our approach in other domains and on other language pairs. Finally, we intend to explore a completely automatic method for short-unit transformation and make it language independent.

References

Ming-Hong Bai, Keh-Jiann Chen, and Jason S. Chang. (2008). Improving word alignment by adjusting Chinese word segmentation. In Proceedings of the Third International Joint Conference on Natural Language Processing, Pp.249-256

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning (2008). Optimizing Chinese word segmentation for machine translation performance. *In Proceedings of the Third Workshop on Statistical Machine Translation*, pp.224-232.

Christopher Dyer, Smaranda Muresan, and PhilipResnik(2008). Generalizing word lattice translation. *Processing of ACL 2008-HLT*, pp.1012-1120

Philipp Koehn, Franz Josef Och, and Daniel Marcu.(2003) Statistical phrase-based translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* , pp.127-133

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa and Hitoshi Isahara (2009). An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. *In Proceedings of the Joint conference of the 47th*

Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP), Singapore, pp.513-521,

Yanjun Ma and Andy Way (2009). Bilingually Motivated Word Segmentation for Statistical Machine Translation. *ACM Transactions on Asian Language Information Processing, Vol. 8, No. 2, Article 7*,

Franz Josef Och. (2003). Minimum error rate training for statistical machine translation. *In Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, pp.160-167.

Hideki Ogura, Hanae Koiso, Yumi Fujiike and Yuu Hara (2009). The specifications of morphological information about "the Balanced Corpus of Contemporary Written Japanese", pp41-71.

F. Peng, X. Huang, D. Schuurmans, and N. Cercone. 2002a. Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp.1-7.

Jia Xu, Richard Zens, and Hermann Ney (2004) Do we need Chinese word segmentation for statistical machinetranslation. *In Proc. of the Third SIGHAN Workshop on Chinese Language Learning*,pp122-128.

Ruiqiang Zhang, Keiji Yasuda and Eiichiro Sumita (2008) Chinese word segmentation and Generalizing word lattice translation. ACM Transactions on Speech and Language Processing, Vol. 5, No. 2, Article 4.

(1) http://www.tokuteicorpus.jp/dist/

(2) http://mecab.sourceforge.net/