Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation

EHARA Terumasa*

* Department of Electronic Systems Engineering,
Tokyo University of Science, Suwa
5000-1, Toyohira, Chino-Shi, Nagano 391-0292, Japan
eharate@rs.suwa.tus.ac.jp

Abstract

Since sentences in patent texts are long, they are difficult to translate by a machine. Although statistical machine translation is one of the major streams of the field, long patent sentences are difficult to translate not using syntactic analysis. We propose the combination of a rule based method and a statistical method. It is a rule based machine translation (RMT) with a statistical based post editor (SPE). The evaluation by the NIST score shows RMT+SPE is more accurate than RMT only. Manual checks, however, show the outputs of RMT+SPE often have strange expressions in the target language. So we propose a new evaluation measure NMG (normalized mean grams). Although NMG is based on n-gram, it counts the number of words in the longest word sequence matches between the test sentence and the target language reference corpus. We use two reference corpora. One is the reference translation only the other is a large scaled target language corpus. In the former case, RMT+SPE wins in the later case, RMT wins.

1. Introduction

Sentences in patent texts are long. Figure 1 shows the frequency distribution of sentence length (characters) for Japanese patent text and Japanese newspaper text. The mean length of Japanese patent sentence[1] is 60 characters and of Japanese news sentence is 38 characters.

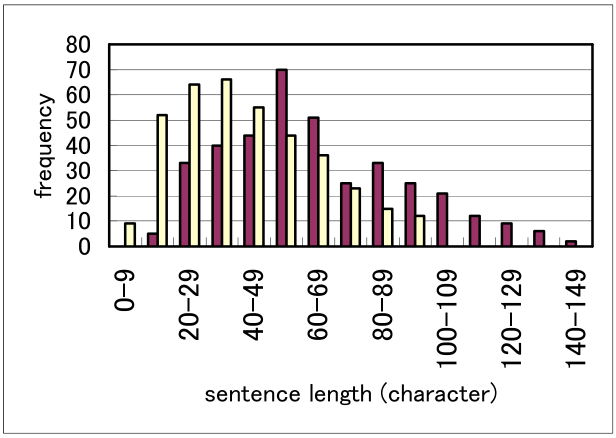


Figure 1: Frequency distribution of the sentence length of Japanese patent text and Japanese news text dark bar: patent; light bar: news

Long sentences are difficult to translate by a machine, because these sentences often have complex syntactic structures. Although statistical machine translation is one of the major streams of the field, long patent sentences are difficult to translate not using syntactic analysis. Some papers show statistical machine translation gives high performance in translation word selection but it often gives syntactically strange outputs. So the combination of a rule based method and a statistical method was one candidate of high quality patent translation. Our system has a structure that combines a rule based machine translation (RMT) with a statistical based post editor (SPE).

There is some research about statistical post processing. (Langkilde and Knight, 1998) uses a statistical post processor in a language generation system. In this system, a symbolic language generator generates the word lattice and a statistical post processor extracts the most appropriate path from the lattice and outputs it. This post processing is controlled by n-gram based language model. (Senef et al, 2006) studies Chinese to English machine translation in the flight domain. They use a SPE system learned from artificially made parallel corpus composed of "bad" English and "good" English sentence pairs. Corpus size is 10,700 sentences. Sentence length is rather short. Mean sentence length of the corpus is 7.3 English words. Recently, (Simard et al, 2007) and (Dugast et al, 2007) used a similar strategy as ours. They are, however, concerning European languages.

In our patent translation case from Japanese to English, we have a parallel corpus. It is "Patent Abstract of Japan (PAJ)" corpus which is manually translated from the abstract part of "unexamined patent publication gazette (PPG)" of Japan[2]. An example of PPG and corresponding PAJ are shown in Appendix 1. So, we can collect "good" English as PAJ sentences and "bad" English as Japanese to English machine-translated results of original Japanese PPG sentences by the RMT.

2. System Architecture

Figure 2 shows the learning process of our statistical post editor. Translation model is learned from PAJ and machine translated results of PPG by RMT. We use GIZA++ as the translation model learner[3]. Language model is learned from PAJ using CMU-Cambridge's language model learner[4]. Figure 3 shows the translation process. Input Japanese patent sentences are translated by

[1] "Problem to be solved" part of "unexamined patent publication gazette of Japan".

[2] http://www.ipdl.inpit.go.jp/homepg_e.ipdl

[3] http://www.fjoch.com/GIZA++.html

[4] http://svr-www.eng.cam.ac.uk/%7Eprc14/toolkit.html

the RMT then they are fed to the SPE. We use the Isi-decoder[5] as the processor of the SPE.

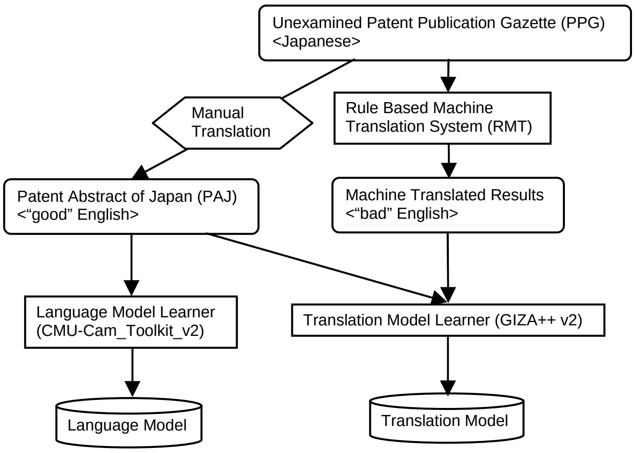


Figure 2: Learning process for the statistical post editor

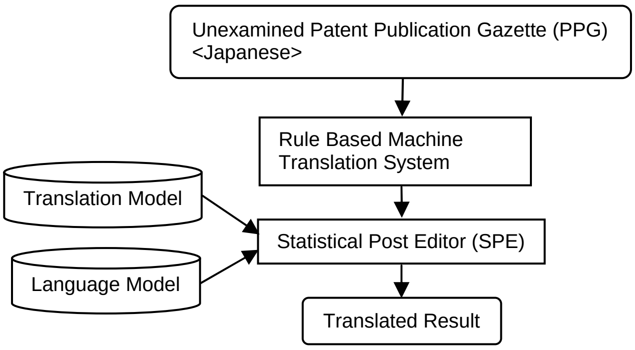


Figure 3: Translation process

3. Translation Experiments

3.1 Training Data and Test Data

We use Japanese and English parallel corpus of patent texts which are described in Chapter 2 as training and test data. They are "unexamined patent publication gazette (PPG)" of Japan as Japanese corpus and corresponding "patent abstract of Japan (PAJ)" as English corpus. We use 2003 year's data. We select only "problem to be solved" part from these corpora, because the first target of our research is to translate this part because it is less complex than the "solution" part.

First of all, we make text alignment between PPG and PAJ using the publication number. Next, we reject aligned texts which have a different number of sentences between PPG and PAJ. Since non-rejected aligned texts have the same number of sentences, we make sentence alignment between PPG and PAJ with the sentence number in the text.

Now, we call the PPG part of sentence aligned corpus as "src" (source sentence) and corresponding PAJ part as "ref" (reference translation). We also call rule based machine translation result of src as "rmt". From this ternary corpus, we make training and test data with the following process:

(1) When the numbers of words of sentences of either rmt or ref are over 90, the datum is rejected.
(2) When the ratio of the numbers of words in sentences of rmt and ref are less than 0.5 or more than 2.0, the datum is rejected.

Through above processes, we get a parallel corpus of src, rmt and ref. From 2003 year's PPG and PAJ original data which includes 337,026 text pairs, we can correct 316,570 sentence ternaries of src, rmt and ref. We use all of them to learn the language model and 92,855 ternaries to learn the translation model. We select 189 ternaries from the set of ternaries which is used for translation model as closed test data and another 189 ternaries from other than this set as open test data.

3.2 Translation Results and Preliminary Evaluation

Using the training data described above, the language model and translation model for SPE are learned. Then the translation system shown in Figure 3 is constructed. We call the output of RMT+SPE system "spe". We do closed and open test using the test data. We compare our results with base-line result that is the output of RMT only, that is "rmt" part of the ternary corpus. Some examples of test results are listed in Appendix 2. For the preliminary evaluation of the translation accuracy, we use the sentence level NIST score which needs reference translation(s). We use "ref" data as the reference. Therefore the number of reference is one. NIST scores are shown in Table 1.

Table 1: NIST scores
μ : mean; σ : standard deviation

test data	system	NIST	
		μ	σ
closed	rmt	4.274	1.329
	spe	5.198	1.769
open	rmt	4.423	1.262
	spe	4.871	1.498

The Kolmogorov-Smirnov test shows that all NIST scores belong to the normal distribution, with significant level 0.05. By the dependent t-test, spe provides significantly accurate translations than rmt, with significant level 0.01, both in closed and open test.

Manual check of the translation results by a human, however, reveals spe results often include syntactically strange expressions than rmt results. We guess that NIST is problematic to measure the translation accuracy, especially, the fluency as the target language. The BLEU case, (Callison-Burch et al., 2006) shows such problems.

3.3 A New Evaluation Measure NMG

To evaluate fluency measure, we need to use not only the small sized reference translation(s) but large sized target language corpus. We use US patent corpus as the target language corpus. Using this large sized reference corpus, we define a new evaluation measure of translation accuracy named NMG as follows.

(1) We consider that the test sentence C is constructed n words: $w_1, \cdots w_n$.

[5] http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html

(2) For each w_i, we define $grams\ (w_i)$ as the maximum number of m that satisfies $w_i, \cdots w_{i+m} \in R$ where R is the set of all n-grams in the reference corpus.

(3) We define NMG score of C as

$$NMG(C) = \log_e(\sum_{i=1}^{n} grams(w_i)/n)$$

For example, if reference corpus includes the following four sentences:

i am a boy
you are a girl
he is a man
she is a woman

and when the test sentence C is

she is a girl

then, n=4 and

grams(she)=3
grams(is)=2
grams(a)=2
grams(girl)=1

Then

$$NMG(C) = \log_e((3+2+2+1)/4) = \log_e(2.00) = 0.69$$

3.4 Evaluation Using NMG

To evaluate RMT and RMT+SPE using NMG, we use two kinds of reference corpus. One is the same as the reference corpus which is used at the NIST score calculation. That is the corpus constructed by only one "ref" sentence which is in the PAJ. This reference corpus is named REF. The other is the corpus including 819,123 sentences extracted from the abstract part of the 157,596 US patent descriptions in the year 2000. This reference corpus is named ABS. We call NMG score using REF as NMG_REF and NMG score using ABS as NMG_ABS. When calculating NMG_ABS, we, however, ignore the following words as the stop words, because of reduction in the index file size.

the, a, of, ",", ".", and, to, is, in, an, for, with, by, which, from, at, on, be

We put the grams value of the above words as zero and, instead, we subtract the number of stop words from the word counts n.

The evaluation results using NMG are listed in Table 2.

Table 2: Evaluation Results using NMG
μ : mean; σ : standard deviation

test data	system	NMG_REF		NMG_ABS	
		μ	σ	μ	σ
closed	rmt	-0.1973	0.3802	0.7777	0.1798
	spe	0.1237	0.4839	0.7449	0.2005
open	rmt	-0.1463	0.3498	0.7795	0.1390
	spe	0.0533	0.3976	0.7159	0.1842

In NMG_REF case, spe wins rmt both in the closed and open test. In NMG_ABS case, rmt wins spe both in the closed and open test. These results suggest that spe has the advantage in "adequacy" and rmt has the advantage in "fluency". The Kolmogorov-Smirnov test shows that all NMG scores belong to the normal distribution, with significant level 0.05. By the dependent t-test, the differences between spe and rmt are significant with significant level 0.01 both in the closed and open test. Figure 4 shows the distribution of the difference of NGM_REF of spe and rmt in the open test. Figure 5 shows the distribution of the difference of NGM_ABS of spe and rmt in the open test.

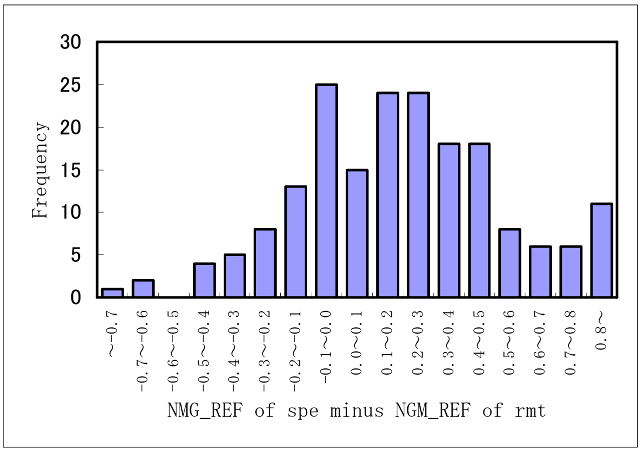


Figure 4: Distribution of the difference of NMG_REF

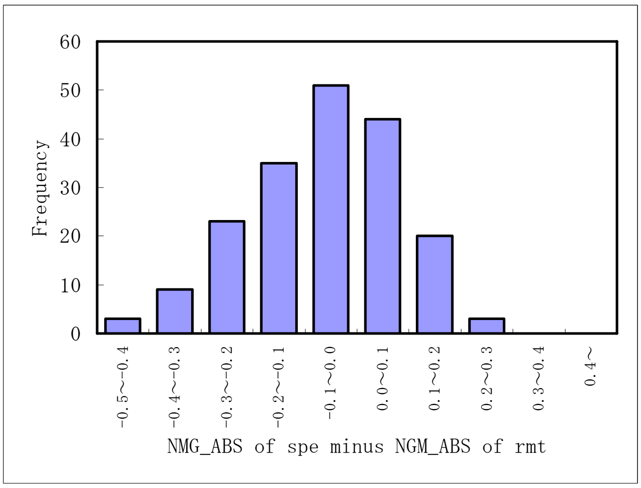


Figure 5: Distribution of the difference of NMG_ABS

3.4 Correlations between NIST and NMG_REF and between NMG_REF and NMG_ABS

Figure 6 shows the correlation between NIST score and NMG_REF score for the closed data. These data come from spe. Pearson's correlation coefficient between NIST and NMG_REF is 0.867. They are highly correlated.

Figure 7 shows the correlation between NMG_REF score and NMG_ABS score for closed test data of the spe system. Pearson's correlation coefficient between NMG_REF and NMG_ABS is 0.356. They are almost uncorrelated.

4. Related Works

Some researchers proposed translation accuracy evaluation measures using a large target language corpus (Callison-Burch & Flournoy, 2001; Akiba et al., 2002; Nomoto, 2003; Quirk, 2004; Corston-Oliver & Gamon, 2001; Kulesza & Shieber, 2004; Gamon et al., 2005). They use n-gram based perplexity type language models and/or syntax/semantic based language models to evaluate translation accuracy. Syntax/semantic based model has the drawback that it needs lots of linguistic knowledge compared with n-gram based model. Our model is also based on n-gram, however, we do not use perplexity but the number of words of longest word sequence match. We do not find such an approach in previous works. (Miyashita et al., 2007) uses sentence match with the web corpus to evaluate fluency of the translation results, but it does not use word sequence match.

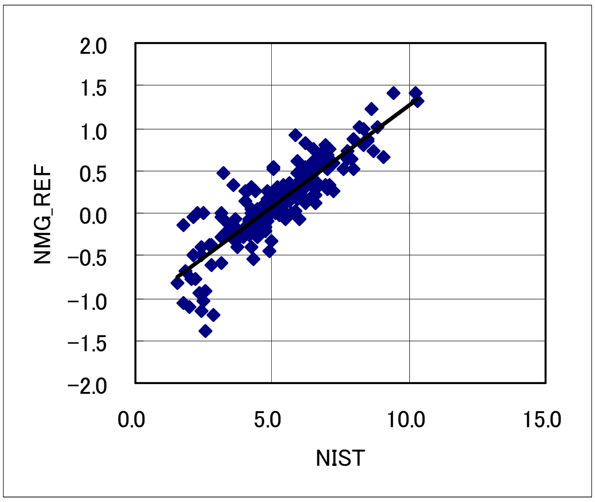


Figure 6: Correlation between NIST and NMG_REF

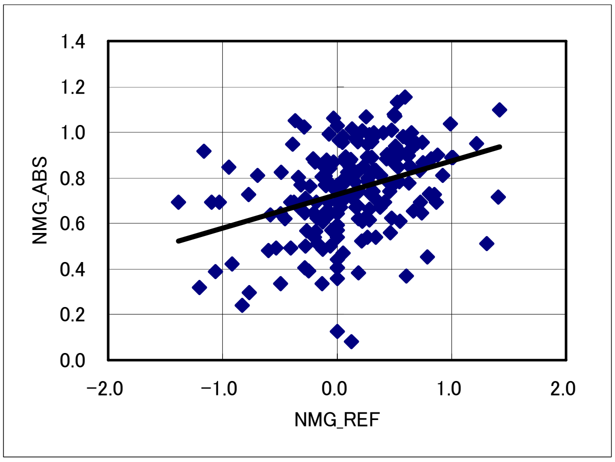


Figure 7:Correlation between NMG_REF and NMG_ABS

5. Conclusion

We proposed a rule based machine translation combined with statistical based post editing. In the evaluation process of our system, we proposed a new n-gram based measure NMG to evaluate translation accuracy. It uses word sequence match with reference translation(s) or large scaled target language corpus. From this evaluation result, we conclude the rule based part of the system has an advantage for structural transfer of a long and complex sentence, which is frequently seen in patent texts. On the other hand, the statistical part of the system has an advantage for lexical transfer of highly technical terms, which is also frequently seen in patent texts.

One of the future works is to compare NMG data to human evaluation results.

Acknowledgements

This work is done under the research in the AAMT/Japio Special Interest Group on Patent Translation. The author expresses sincere acknowledgements to the group members for their useful discussions. The corpora used in the work are provided by Japan Patent Information Organization (Japio). The author sincerely acknowledges Japio for their support.

References

Yasuhiro Akiba; Taro Watanabe and Eiichiro Sumita (2002): Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems, COLING2002.

Chris Callison-Burch and Raymond S. Flournoy (2001): A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines, MT Summit VIII, 2001.

Chris Callison-Burch; Miles Osborne and Philipp Koehn (2006): Re-evaluating the Role of BLEU in Machine Translation Research, EACL, 2006.

Simon Corston-Oliver; Michael Gamon and Chris Brockett (2001): A Machine Learning Approach to the Automatic Evaluation of Machine Translation, ACL2001.

Loïc Dugast; Jean Senellart and Philipp Koehn (2007): Statistical Post-Editing on SYSTRAN's Rule-Based Translation System, Proc. of the Second Workshop on Statistical Machine Translation, pp.220-223.

Michael Gamon; Anthony Aue and Martine Smets (2005): Sentence-level MT Evaluation without Reference Translations: Beyond language modeling, EAMT2005.

Alex Kulesza and Stuart M. Shieber (2004): A Learning Approach to Improving Sentence-Level MT Evaluation, TMI2004.

Irene Langkilde and Kevin Knight (1998): Generation that Exploits Corpus-Based Statistical Knowledge, ACL/COLING1998.

Kohei Miyashita; Seiichi Yamamoto; Keiji Yasuda and Masuzo Yanagida (2007): Quality Evaluation Method of Machine Translated Sentences by Comparing Text Retrieved from Web and Using Translation Model, Information Processing Society of Japan Special Interest Group Technical Reports, NL-177, pp.17-23, 2007 (in Japanese).

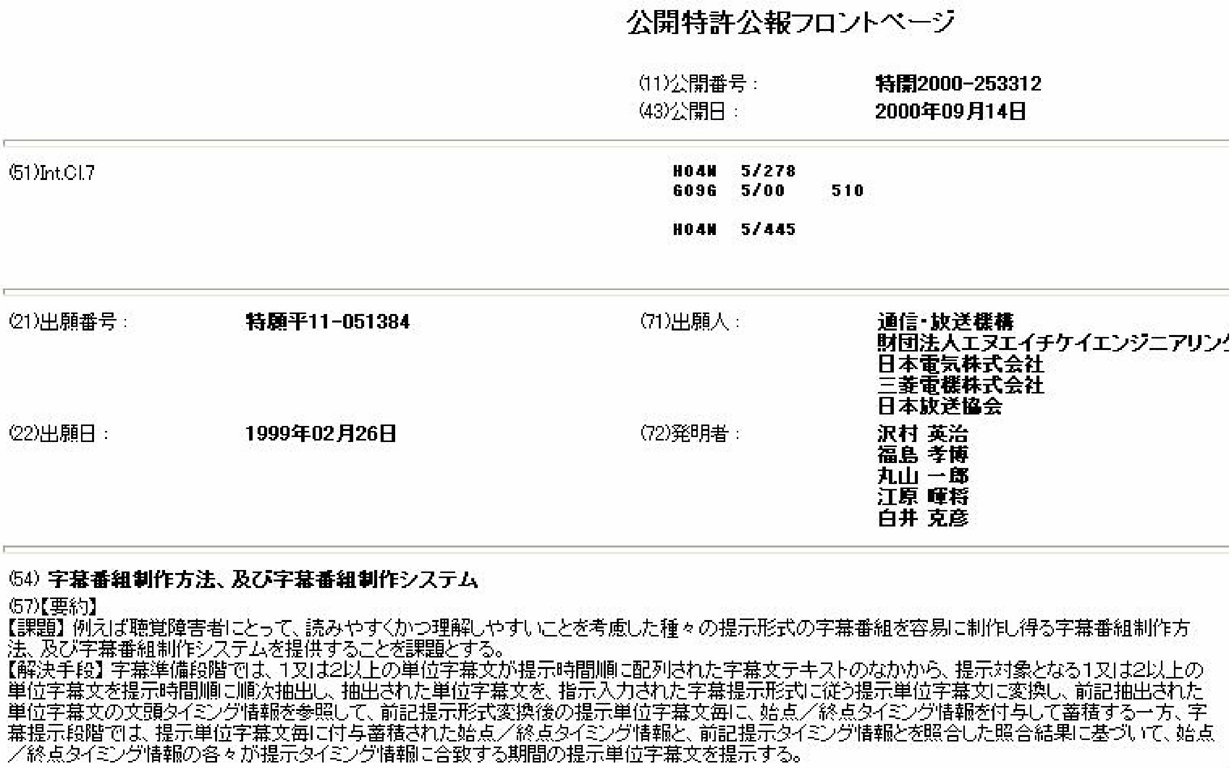
Tadashi Nomoto (2003): Predictive Models of Performance in Multi-Engine Machine Translation, MT Summit IX, 2003.

Christopher B. Quirk (2004): Training a Sentence-Level Machine Translation Confidence Measure, LREC2004.

Stephanie Seneff; Chao Wang and John Lee (2006): Combining Linguistic and Statistical Methods for Bi-directional English Chinese Translation in the Flight Domain, AMTA2006.

Michel Simard et al. (2007): Rule-based Translation with Statistical Phrase-based Post-editing, Proc. of the Second Workshop on Statistical Machine Translation, pp.203-206.

Appendix 1 Unexamined patent publication gazette and corresponding patent abstract of Japan[6]



公開特許公報フロントページ

(11)公開番号： 特開2000-253312
(43)公開日： 2000年09月14日

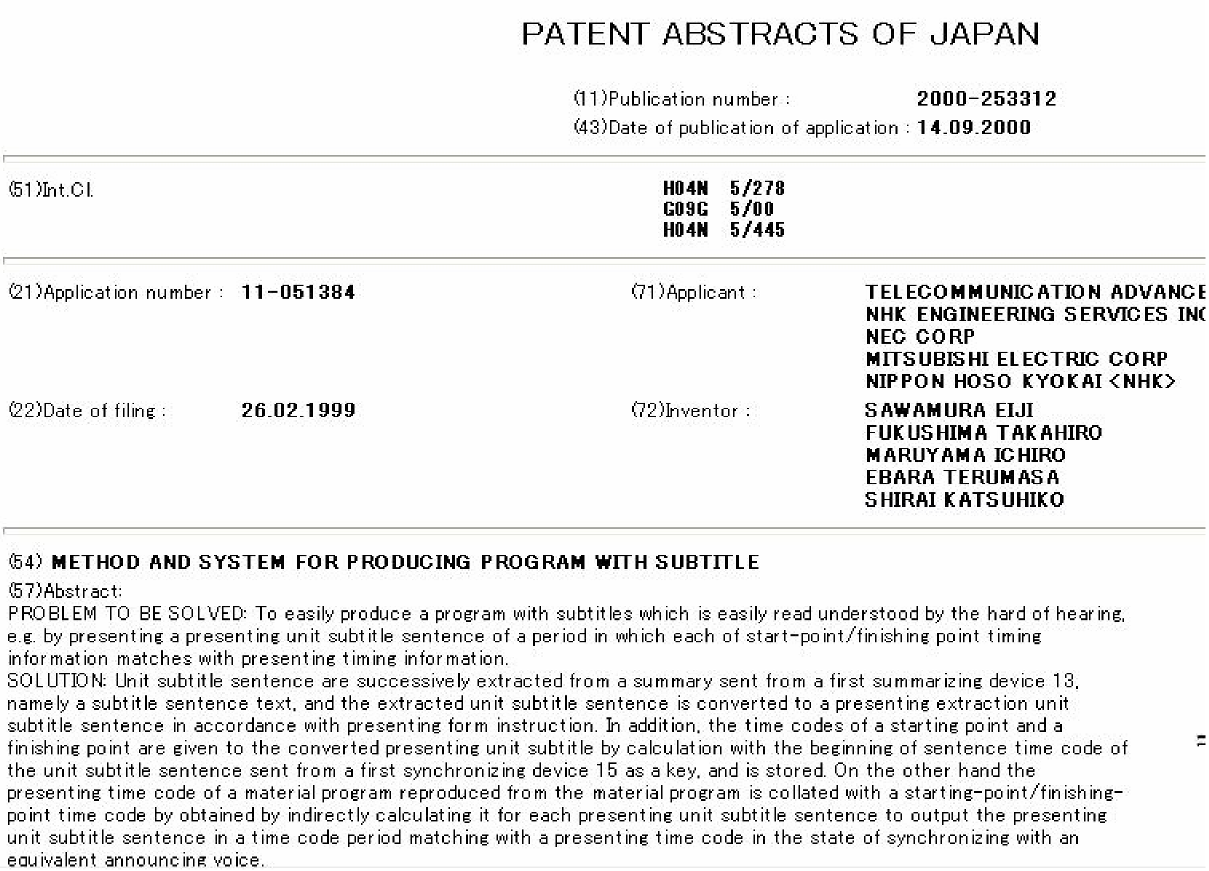
(51)Int.Cl.7
H04N 5/278
G09G 5/00 510
H04N 5/445

(21)出願番号： 特願平11-051384
(71)出願人： 通信・放送機構
財団法人エヌエイチケイエンジニアリン
日本電気株式会社
三菱電機株式会社
日本放送協会

(22)出願日： 1999年02月26日
(72)発明者： 沢村 英治
福島 孝博
丸山 一郎
江原 暉将
白井 克彦

(54) **字幕番組制作方法、及び字幕番組制作システム**

(57)【要約】
【課題】例えば聴覚障害者にとって、読みやすくかつ理解しやすいことを考慮した種々の提示形式の字幕番組を容易に制作し得る字幕番組制作方法、及び字幕番組制作システムを提供することを課題とする。
【解決手段】字幕準備段階では、1又は2以上の単位字幕文が提示時間順に配列された字幕文テキストのなかから、提示対象となる1又は2以上の単位字幕文を提示時間順に順次抽出し、抽出された単位字幕文を、指示入力された字幕提示形式に従う提示単位字幕文に変換し、前記抽出された単位字幕文の文頭タイミング情報を参照して、前記提示形式変換後の提示単位字幕文毎に、始点／終点タイミング情報を付与して蓄積する一方、字幕提示段階では、提示単位字幕文毎に付与蓄積された始点／終点タイミング情報と、前記提示タイミング情報とを照合した照合結果に基づいて、始点／終点タイミング情報の各々が提示タイミング情報に合致する期間の提示単位字幕文を提示する。



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-253312
(43)Date of publication of application : 14.09.2000

(51)Int.Cl.
H04N 5/278
G09G 5/00
H04N 5/445

(21)Application number : 11-051384
(71)Applicant : TELECOMMUNICATION ADVANCE
NHK ENGINEERING SERVICES IN
NEC CORP
MITSUBISHI ELECTRIC CORP
NIPPON HOSO KYOKAI <NHK>

(22)Date of filing : 26.02.1999
(72)Inventor : SAWAMURA EIJI
FUKUSHIMA TAKAHIRO
MARUYAMA ICHIRO
EBARA TERUMASA
SHIRAI KATSUHIKO

(54) **METHOD AND SYSTEM FOR PRODUCING PROGRAM WITH SUBTITLE**

(57)Abstract:
PROBLEM TO BE SOLVED: To easily produce a program with subtitles which is easily read understood by the hard of hearing, e.g. by presenting a presenting unit subtitle sentence of a period in which each of start-point/finishing point timing information matches with presenting timing information.
SOLUTION: Unit subtitle sentence are successively extracted from a summary sent from a first summarizing device 13, namely a subtitle sentence text, and the extracted unit subtitle sentence is converted to a presenting extraction unit subtitle sentence in accordance with presenting form instruction. In addition, the time codes of a starting point and a finishing point are given to the converted presenting unit subtitle by calculation with the beginning of sentence time code of the unit subtitle sentence sent from a first synchronizing device 15 as a key, and is stored. On the other hand the presenting time code of a material program reproduced from the material program is collated with a starting-point/finishing-point time code by obtained by indirectly calculating it for each presenting unit subtitle sentence to output the presenting unit subtitle sentence in a time code period matching with a presenting time code in the state of synchronizing with an equivalent announcing voice.

[6] Quoted from Industrial Property Digital Library of Japan: http://www.ipdl.inpit.go.jp/homepg_e.ipdl

Appendix 2 Translation Examples and their NIST and NMG socres

closed test 2003000001

[src]

帯状土塊を破砕するとともに、培土作業を停止すると抱き込まれていた土塊が自動的に落下する構造の培土器を提供する

[ref]

to provide a ridger having a structure capable of crushing zonal clod and automatically dropping caught clod when ridging operation is stopped .

[rmt]

to provide structural Tstica that it wins to over one's side and the clod falls automatically when a zonal clod is crushed , and the Tstica work is stopped .

[spe]

to provide a ridger which is ridger to side and a clod falls down automati when a band-shaped screened is ignitionability and a drip-watering operation is stopped

	rmt	spe
NIST	3.4828	4.3577
NMG_REF	-0.4054	0.0000
NMG_ABS	0.3136	0.1251

open test 2003102102

[src]

電動機や制御回路が過負荷を受けるのを未然に防止するとともに、スムーズな乗り心地を実現させることのできる電気自動車の制御装置を提供する。

[ref]

to provide a control device for an electric vehicle , which prevents overload on an electric motor and a control circuit , and to realize smooth ride .

[rmt]

to prevent the electric motor and the controlling circuit from receiving the overload beforehand , and to provide the controller of the electric vehicle that can achieve smooth riding comfort .

[spe]

to prevent a electric motor and a control circuit from receiving an over load in advance , and to provide a transmission) of a motor-driven vehicle which can realize planarizing riding comfortableness

	rmt	spe
NIST	4.0614	4.8926
NMG_REF	-0.3429	0.2231
NMG_ABS	0.8708	0.7985