Automatic extraction of differences between spoken and written languages, and automatic translation from the written to the spoken language

Masaki Murata and Hitoshi Isahara

Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{murata,isahara}@crl.go.jp

Abstract

We extracted the differences between spoken language and written language from a spoken-language corpus and a written-language corpus by using the UNIX command “diff” and examined the differences to determine the construction of the grammars of the two corpora. We also transformed written-language sentences into spoken-language sentences by using rules based on the extracted differences.

1. Introduction

The Communications Research Laboratory and the National Institute for Japanese Language have constructed a corpus of the Japanese spoken language. The corpus consists of data gathered from presentations at academic conferences. In the corpus, the spoken data (i.e., the presentations) was transcribed and stored as text data. We then typed the papers corresponding to those presentations, so we now have the spoken-language text data (transcriptions of the presentations) and the written-language text data (the papers) together.

In this study, we used those data for the following two purposes.

- Automatic extraction of differences between spoken and written languages:

  We detected the differences between spoken and written languages by matching the spoken data and the written data by using the UNIX “diff” command. We examined the differences to clarify the distinctions and the similarities between spoken and written languages.

- Automatic translation from the written language to the spoken language:

  The detected differences can be considered to be the transformation rules between spoken and written languages. In this work, we actually use the differences as the transformation rules and transform a written text to a spoken-language-like text.[1]

Table 1: Sample data of written and spoken language

Written language	Spoken language
In	Today
this	I’d
paper,	like
we	to
describe	describe
the	uh
meaning	the
sort.	meaning
In	sort.
general,	In
sorting	general,
is	sorting
performed	is
by	done
using	by

2. How to detect differences between spoken and written languages

In this section, we explain our method of detecting differences between spoken and written languages.

We first divided the spoken and written data into words by using a Japanese morphological analyzer. The results are shown in Table 1. (In this sample text, we used English, not Japanese, for English readers.)

We used the UNIX command, “diff”, to determine the differences between the spoken and written data and obtained the results shown in Table 2. (“;===== begin =====” indicates the beginning of a difference and “;===== end =====” indicates the end of a difference. “;—————–” indicates the boundary of two parts of a difference.)

When we extracted the different parts, we obtained the results shown in Table 3.

The results show that “uh” is inserted in the spoken language and that “performed” can, in this case, be paraphrased as “done.” In this sample data, we could easily see the differences between the spoken and written language. However, in actual papers and presen-

[1] In this paper, we transformed the written language to the spoken language. However, we will be able to transform the spoken language to the written language in the same way. We have already constructed the universal model for paraphrasing (Murata and Isahara, 2001b). In this model, we could construct various types of paraphrasing systems including one for answering questions, one for compressing sentences, one for polishing up, and one for transforming written language to spoken language.

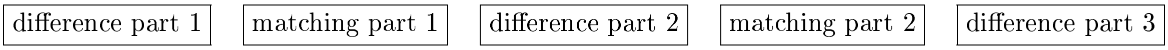

Figure 1: Expansion of difference parts

Table 2: Result of “diff” of written and spoken data

```
;===== begin =====
In
this
paper,
we
;——————
Today
I'd
like
to
;===== end =====
describe
;===== begin =====
;——————
uh
;===== end =====
the
meaning
sort.
In
general,
sorting
is
;===== begin =====
performed
;——————
done
;===== end =====
by
```

Table 3: Extraction of differences

written data	spoken data
“In this paper, we”	“Today I'd like to”
“”	“uh”
“performed”	“done”

tations, we often say very different things. Therefore detecting differences becomes more difficult. We made a more accurate method for detecting differences. The method uses probablistic equations using the following two characteristics:

- The better differences are surrounded by lower-frequency words.
- The better differences occur more frequently.

(“The better differences” mean the differences that are more useful for linguistic investigation on the differences between the spoken and written language and for the transformation rules between spoken and written languages.)

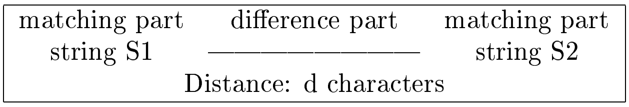

Figure 2: Occurrence of differences

First, we considered the first characteristic, “the better differences are surrounded by lower-frequency words.” We assumed a difference part is surrounded by strings, $S1$ and $S2$, which are matching parts and the distance between $S1$ and $S2$ is d characters[2] as in Figure 2. The probability, $P(S1)$ or $P(S2)$, of occurrence of $S1$ or $S2$ in the inner region consisting of no more than d characters from $S2$ or $S1$ are approximately expressed as follows:

$$P(S1) \simeq (d+1) * \frac{Freq(S1)}{N} \quad (1)$$

$$P(S2) \simeq (d+1) * \frac{Freq(S2)}{N} \quad (2)$$

where $Freq(S1)$ and $Freq(S2)$ are the numbers of occurrence of strings $S1$ and $S2$, and N is the total number of characters in the database. When we assumed the probability $P(dfp, S1, S2)$ that the difference part (dfp) is good is the probability that strings $S1$ and $S2$ do not appear in the situation shown in Figure 2, $P(dfp, S1, S2)$ is expressed as the following:

$$P(dfp, S1, S2) \quad \simeq \quad (1 - P(S1))(1 - P(S2)) \quad (3)$$

where we assume that $S1$ and $S2$ are independent of each other.

Next, we considered the second characteristic, “The better differences occur more frequently.” We have only to combine the probabilities in plural situations. We assumed that when at least one of the plural situations is correct, we extract the difference part as a correct one. Since the fact that the difference part is correct is the complement of the case when all the situations for the difference part are incorrect, the probability $P(dfp)$ that the difference part (dfp) is correct is expressed as the following equation.

$$P(dfp) \quad \simeq \quad 1 - \prod_{S1,S2} (1 - P(dfp, S1, S2)), \quad (4)$$

where we assume that each situation of difference parts is independent of one another.

[2] In this study, we use a longer length of characters of differences as d.

Table 4: Number of extracted differences

	Number of extracted differences
Probability ≥ 99.99%	1,011
Probability ≥ 99.9%	3,245
Probability ≥ 99%	7,846
Probability ≥ 90%	14,777
Total	72,835
Frequency ≥ 2	421

The extraction of difference parts are performed by sorting the results of "diff" by the value of the above equation and extracting the one having a higher value.

We made the following modification to the process used to extract the candidates of difference parts.

- When matching parts and difference parts appear as in Figure 1, we added "difference part 1, matching part 1, difference part 2" and "difference part 1, matching part 1, difference part 2, matching part 2, difference part 3" to the list of candidates of difference parts.

In this study, we restricted the difference parts generated by this connection to the ones including at the maximum three original difference parts.

3. Extraction and examination of differences

In this section, we describe the experiments performed to extract differences using our method described in the previous section.

We used, as written and spoken data, 82 electronic data (papers and presentations at academic conferences) which had been constructed.

- Written-language data
  - Papers (82 papers, 352,660 characters)
- Spoken-language data
  - Presentations
    The presentations corresponding to the above papers (330,679 characters)

We extracted differences between written and spoken data using the method described in Section 2. The number of extracted differences is shown in Table 4. Probabilities in the table mean the values of Equation (4). The top 40 extracted differences were sorted according to the value given by Equation (4) in Section 2. Frequency in the table means the number of occurrence of differences.

We examined the results of detected differences and found the following:

1. Inconsistency of expressions

   Examples of inconsistent expressions are shown in Table 6. "data" can be expressed as "deeta" or "deetaa" in Japanese. "all" can be expressed as "subete (hiragana character)" or "subete (kanji, chinese character)" in Japanese. These are inconsistent expressions. Such inconsistencies were extracted by our method.

2. Showing how to read

   We show some examples for this in Table 7. The first line indicates "=" (equal) in the written data is expressed as "wa" (be) in the spoken data. From this, we can find that "=" (equal) was said such as "wa" (be) in Japanese. These examples showed how to read Japanese written expressions.

3. Synonyms

   We show some examples for extracted synonyms in Table 8. Since we used expressions in presentations that are different from those written in papers, we can extract paraphrases that have the same meaning, as shown in the table. Since the data used in this study were a paper and the presentation corresponding to it, synonyms related to research were extracted.[3]

Table 6: Example of inconsistency of expressions

Written data	Spoken data
deeta	deetaa
(data)	(data)
kurasuta	kurasutaa
(cluster)	(cluster)
kakaru (kanji)	kakaru (different kanji)
(modified)	(modified)
koeru (kanji)	koeru (different kanji)
(exceed)	(exceed)
subete (hiragana)	subete (kanji)
(all)	(all)
tame	tame (kanji)
(for)	(for)
okona-u	oko-nau
(perform)	(perform)
iikae (kanji)	iikae (different kanji)
(paraphrase)	(paraphrase)

Table 7: Example of showing how to read

Written data	Spoken data
=	wa
(equal)	(be)
2	ni (kanji)
(two)	(two)
rei	zero
(zero)	(zero)
-gram	guramu
(-gram)	(gram)
s	byou
(s)	(second)
Hebb	hebu
(hebb)	(hebb)

[3] A study of extraction of paraphrases is important in

Table 5: Example of results of matching written and spoken data (Top 40)

Freq.	examples of matching part in front	written	spoken	examples of matching part behind
182	IPAL no keiyoushi	,		keiyou-doushi no
72	wo bokokugo		no (of)	kiiwaado de kouritsu yoku kensaku
43	pata pata	.		bata bata
49	no kakutyou de aru		e (eh-)	shikibetsu teki tokutyou
56	LR hyou he no hukusuu no		ee (eh-)	setsuzoku seiyaku
54	hyaku man en ni naru	”		to yogen shita toka
39	ni zokusuru bekutoru	no (of)		wa no kyori ni
28	honbun kan no haipaarinku		wo (obj.)	jidou seisei
19	shuushoku youso	wo (obj.)		torikomu
21	shiborikomi go wa	“		bei eiga
22	meishi no kurikaeshi no baai	,	ee (eh-)	sentou no meishi nomi
20	wai wa	)		x to jijou ga chigau
21	son shitsu wa zen gakushuu	deeta (data)	deetaa (data)	ni taisuru sonshitsu no
11	oyobi yougen no	ga (katakana) (sub.)	ga (hiragana) (sub.)	kaku jouhou wo huyo
19	gakkai happyou ronbun		<C>	25534 hyoudai tsui kara naru
13	tairyou no koupasu wo mochiite	,	e (eh-)	kikai hon-yaku ni yori
15	mado kansuu		ni (by)	yori kiridasareta
12	gengengo to mokuhyou gengo	to (and)		no aida no douji shinkou sei ga
10	shimesu kotoba wo oginatte	1 (one)	ichi (kanji) (one)	tsu no bun wo
10	zatsuon supekutoru	no ('s)	wo (obj.)	genzan
14	sono kekka wo hitode		de (by)	shuusei shite iku
20	sukunai		toiu (that is)	koto kara
10	bangumi	wo (obj.)	no (of)	saisho kara saigo made
10	goukei kyoudo ga N	_{		ε
11	kiji no		sono (its)	kouzou to tokutyou
16	bangumi jidou jimaku ka no	tame (hiragana) (for)	tame (kanji) (for)	no onsei ninshiki shisutemu wo
10	wari chikai	kurasuta (cluster)	kurasutaa (cluster)	ga imi
6	bekutoru wo kotonaru	k	K	ko no shuugou ni
8	VQ koudo bukku no	2 (two)	ni (kanji) (two)	shurui no washa moderu
7	tekigou ritsu	= (equal)	wa (be)	honyaku kekka ga
8	sono buntyuu de		n (“nn”)	wadai to natte iru youso
9	juubun yoi seido de		wa (be)	suitei dekinai toiu mondai ga
6	renketsu gakushuu to		o (“oh”)	kongou suu wo baizou suru
11	picchi no joushou		toiu mono (that is)	ga yuusei on
7	kiso teki		na (-tive)	kentou wo
5	seiki seigen	no (of)	ga (sub.)	kibishii keitaiso mo
5	kyou taiiki	/		kou taiiki CELP
7	hukas D no		oo (“oh”)	joui gainen ni tyuushou ka
7	wo kontekisuto ni motsu		youna (such that)	gengo moderuni
14	shuhou		toiuno (that is)	wo teian

Table 10: Example of ellipsis

Matching part in front	Written data	Spoken data	Matching part behind
sumuujingu (smoothing)	shori (process)		wo (obj. case-particle)
kaku (each)	C(V)_{k}		sohen (piece)
supoutsu nyuusu (sports news)	ni okeru (in)	no (of)	kaiwa bubun wo (conversation)
heikin jikan ga (average time)	11.25 (11.25)	11.3 (11.3)	hun made (minutes)

Table 11: Example of complementation

Matching part in front	Written data	Spoken data	Matching part behind
sonshitsu no (loss)		atai no (values)	heikin to shite (in average)
kaiwa ni (conversation)		kanshi mashitewa zenzen (about)	hugen wa nai (no inconvenience)
on-atsu reberu (sound level)	70dB (70 dB)	nanajuu gogo deshiberu (70.55 dB)	de teiji (shown)

Table 8: Example of synonyms

Written data	Spoken data
oyobi (and)	to (and)
ya (or)	toka (or)
ronbun (paper)	kenkyu (study)
,kotonari (differences)	kotonari-de (differences)
kaku (each)	sorezore (each)
i-banme no taamu (i-th term)	taamu I (term I)
jutsugo (predicate)	doushi (verb)
shikibetsu (discrimination)	ninshiki (recognition)
kotonareba, (different)	chigaeba (different)

Table 9: Example of colloquial style

Written data	Spoken data
	toiu (that is)
shita. (did)	itashi mashita (did, polite expression)
	desu (polite expression)
rareru. (be done)	raremasu (be done, polite exp.)
	tteiu (that is)
	kou (this)

4. Colloquial style

   We show some examples of this style in Table 9. We extracted many colloquial-style Japanese expressions. “toiu”, which is “that is” in English, is extracted. “toiu” is a colloquial expression. Such an expression was extracted by matching written and spoken data. Both “shita” and “itashi mashita” mean “did”. However, “itashi mashita” is much more polite than “shita”. In Japanese, we use polite expressions in presentations and we use neutral expressions in papers. Hence, we extracted the pairs by matching written and spoken data. “kou” (this) was extracted. In spoken language, the denotation expressions such as “this” are often used. “kou” would be one example of them.

5. Ellipsis (decrease of information)

   We show some examples of ellipsis in Table 10. “process” was omitted in the spoken data. “11.25” in the written data was changed into the lighter expression “11.3” in the spoken data.

6. Complementation (increase of information)

   We show some examples for complementation in Table 11. This is the inverse of the above item “Ellipsis”. “loss in average” in written data was changed into a richer expression “the value of loss in average”. “70 dB” in written data was also changed into a richer expression “70.55 dB.”

natural language processing. We used the method described in this study for different dictionaries and obtained many synonyms (Murata and Isahara, 2001b).

Table 12: Examples of error detection

Matching part in front	Written data	Spoken data	Matching part behind
Errors ocurr in written data			
nyuusu	sokki	sokuhou	kiji
(news)	(shorthand)	(prompt report)	(articles)
nihongo ga	nobe (kanji)	nobe (different kanji)	178091 go
(Japanese)	(describe)	(total number)	(178091 words)
nihongo hukugou	meishi	meishi kouzou	kaiseki hou
(Japanese compound)	(name card)	(noun structure)	(analysis method)
Errors occur in spoken data			
kagi kakko no shurui wo	hyousou tekina	hyousyoutekina	jouhou nomi de
(sorts of parenthesis)	(surface)	(binding)	(information)
One of both will be incorrect			
maikuro hon	oyobi seitai (kanji)	oyobi seitai (different kanji)	an-pu
(microphone)	(and organism)	(and vocal cords)	(amplifier)
shakai no	shikatsu	seikatsu	ni kakawaru mondai
(social)	(fate)	(life)	(problem)

In general, spoken language has more elliptical expressions than written language. These cases, such as complementation, are counter examples of them, therefore they are interesting.

7. Error detection

   We show some examples of error detection in Table 12. When a written data and a spoken data included errors, such errors were occasionally extracted as differences. The first line of the table indicates that "short hand" was a wrong expression and "prompt report" was a correct expression. This error would be made when a paper was transformed into an electronic data. We also found errors in the spoken data.

In this section, we extract differences between a written language and a spoken language and examine the results. Our study, which used computational processing for a study of the differences between written and spoken languages, would be an interesting approach.

4. Trial of automatic translation from the written to the spoken language

We extracted many differences between written and spoken languages in the experiments described in the previous section. We can regard the extracted differences as transformation rules from written to spoken languages. In this section, we handled the detected differences as transformation rules and tried to automatically transform written language into spoken language.

The transformation rules used were the top 240 differences in the list sorted according to the value of Equation (4). The differences after the top 240 included differences whose frequency was one. We judged that a difference whose frequency is one was not reliable. Thus, we did not use the differences below the top 240.

We used the following procedures for transforming the written language into the spoken language.

1. The system analyzed an input sentence morphologically by using the Japanese morphological analyzer JUMAN (Kurohashi and Nagao, 1998) and divided it into a string of morphemes.

2. The system performed the following procedures for each morpheme from left to right.

   (a) When the string of morphemes S, whose first morpheme is the current one (including no morphemes, e.g., "") that matched the A_i string from the transformation rule R_i ($A_i \Rightarrow B_i$), the B_i string was extracted as the candidate of the transformed expression. We referred to the string of the k-gram morphemes just before S as $S1_i$ and to the one just after as $S2_i$.

   (b) The system counted the number of the frequency of $S1_iB_iS2_i$ when string A_i was changed to B_i against each B_i. We referred to the i when the value was the highest as m.

   (c) The system counts the frequencies of the strings of $S1_mA_mS2_m$ and $S1_mB_mS2_m$ in the corpus. When the number of the frequency of $S1_mB_mS2_m$ exceeds that of $S1_mA_mS2_m$, the system transforms A_m to B_m and performs the procedure of the next morphology.

where, k is a constant.

Roughly speaking, this algorithm transforms sentences so that each expression in the sentences have higher frequencies in the spoken language data, that is, this algorithm transforms an expression into the expression occurring frequently in spoken data.[4]

[4] To improve the results of transformation, the frequency of each string x in our procedures must be changed to the

Table 13: Examples of transformation from written to spoken language (the case of k=1)

	kin-nen	chishiki		kakutoku	no	kenkyu	ga juuyoushi sare tsutsu aru.
	(recently)	(knowledge)		(extraction)	(of)	(study)	(became important)
e			wo				
(eh)			(obj.)				
(Eh, study on knowledge (obj.) extraction became important recently.)							

hon	kou	dewa	,	dougi no	tekisuto wo	shougou shi	,
(this)	(paper)	(in)	(,)	(same meaning)	(texts)	(match)	(,)
	kenkyu						
	(study)						
(In this paper (⇒ study), we matched texts having the same meaning,)							

sono	shougou	kekka	wo mochiite	chishiki	wo kakutoku shita.
(the)	(matching)	(result)	(use)	(knowledge)	(extracted)
and extracted knowledge using the matching (results).)					

Table 14: Examples of transformation from written to spoken language (the case of k=2)

sono	teigi	wo	riyou	suru	toiu koto	ga		kangae rareru
(its)	(definition)	obj	(use)	(do)	(that)	obj		(think)
							ma	
							filler	
(We can think (ma) that we use its definition.)								

dougi	hyougen	wo	tyuushutsu	suru		koto	wo	kokoromiru.
(same-meaning)	(expression)	obj	(extract)	(do)		(that)	obj	(try)
					toiu			
					(that is/such)			
(We try (such) that we extract the same-meaning expressions.)								

hindo	de	souto	shita	kekka			wo	hyou	ni	shimesu.
(frequency)	(by)	(sort)	(did)	(result)			obj	(table)	(in)	(show)
					toiu	no				
					(that is)	(those)				
(We show (those that is) the results sorted by frequency in the table.)										

We used sentences in our Japanese paper (Murata and Isahara, 2001a) as input and transformed the written language to spoken language by using the above procedure. We showed the result when $k = 1$ in Table 13 and the result when $k = 2$ in Table 14. The underlined part is the part that was removed in the transformation. The lower strings are the transformed ones. Since the algorithm was not so strong, in $k = 1$ the context was short and the precision was low. ("wo" was inserted, but it was the wrong transformation. There were many other wrong transformations in the experiments.) However, good spoken-language-like results were obtained such that "e" (eh) was inserted and "this paper" was transformed into "this study". In $k = 2$, the precision was very good and there were few errors. "toiu" and "ma" are Japanese colloquial expressions and were inserted. The results were very good. However, the number of transformed expressions was very small and the recall rate was very small. We feel obliged to improve the method by using the method described in Footnote 4. We will improve our system by changing the calculation of frequencies into the calculation of probabilities and making the information used for a context richer in our future work.

5. Conclusion

In this study, we extracted differences between spoken and written languages and examined the extracted differences by using spoken and written data constructed by the Communications Research Laboratory and the National Institute for Japanese Language. We also tried transforming written language into spoken language by using extracted differences as the transformation rules.

probability of occurrence of x in the corpora when the given input data is used as the context. Although our procedures use the fixed $k*2$ morphemes of "in front" and "behind" as the context, we must calculate the probabilities by using the variable-length context and more global information, such as syntactic information and tense information, in a powerful probability-estimator such as the maximum entropy method.

Although our examinations of the differences between spoken and written languages were unsufficient, our approach using computational processing for studies of differences between spoken and written languages has demonstrated its efficacy.

We also constructed a basic system for transforming a written language into a spoken language. This system outputted spoken-language-like expressions. The results of our study will be further applied to our future work.

6. References

Sadao Kurohashi and Makoto Nagao, 1998. Japanese Morphological Analysis System JUMAN version 3.5. Department of Informatics, Kyoto University. (in Japanese).

Masaki Murata and Hitoshi Isahara. 2001a. Automatic paraphrase acquisition based on matching of two texts with the same meaning. Information Processing Society of Japan, WGNL 142-18. (in Japanese).

Masaki Murata and Hitoshi Isahara. 2001b. Universal model for paraphrasing — using transformation based on a defined criteria —. In NLPRS'2001 Workshop on Automatic Paraphrasing: Theories and Applications.