Dataset Construction Method for Word Reading Disambiguation

Koki Nishiyama **Kazuhide Yamamoto**
Department of Electrical Engineering
Nagaoka University of Technology
1603-1 Kami-Tomioka, Nagaoka,
Niigata, 940-2188, JAPAN
nishiyama@jnlp.org

Hideharu Nakajima
NTT Communication Science Labs.
NTT Corporation
2-4 Hikari-dai Seika-cho, Soraku-gun,
Kyoto, 619-0237, JAPAN
nakajima.hideharu@lab.ntt.co.jp

Abstract

The scarcity of large corpora in reading disambiguated words is a major limitation in linguistic analysis and the initiation of a statistical approach to word reading disambiguation. As readings of words are usually not written in documents like meanings of words, therefore, human annotation is necessary but expensive. In this study, a method is proposed to construct a reading disambiguated dataset for word reading disambiguation. The method constructs a dataset of sentences wherein words with ambiguity in reading (pronunciation), called heteronyms, are tagged for correct reading. In this method, a word with unique reading is labeled to a heteronym, and this unique word is used as a query word to collect sentences that include the word. The word in the collected sentences is replaced by the original ambiguous word and the reading corresponding to that of the query word is tagged as the pronunciation of the heteronym. It was confirmed through experiments that the method was able to collect data effectively, and the collected data was numerically balanced among all the readings of the heteronym.

1 Introduction

Text-to-speech (TTS) synthesis requires prosodic phrase boundaries and pronunciation (reading and stress/accents) (Hall and Sproat, 2013; Sproat and Hall, 2014). These are determined on the basis of linguistic analysis with natural language processing technologies. Although an adjacent word information such as a statistical language model is applied for the processing, a few words in narrow contexts are used that sometimes lead to errors. Among the errors, those in pronunciation are noticeable in the TTS situations. Furthermore, it has been reported that the subjective evaluation of TTS by listeners is particularly sensitive to homograph disambiguation errors (Braga et al., 2007; Gorman et al., 2018). Thus, homograph disambiguation has been one of the important tasks in synthesis and has been tackled by many researchers in the past (Yarowsky, 1997; Braga et al., 2007).

Homographs are words that share the same written form as other words but have a different meaning, and can be classified into either homonyms or heteronyms, depending on their pronunciation, e.g., homonyms have the same pronunciation, such as *lie* (untruth) [lάi] and *lie* (to recline) [lάi], whereas heteronyms have different pronunciation, such as *desert* (region) [dézərt] and *desert* (to leave) [diz'ə:rt]. Thus, disambiguation of heteronyms is a more important task in TTS synthesis, as the correct pronunciation of the word in its context must be determined.

Statistical methods have been considered to be better techniques for disambiguation, such as word sense/homograph disambiguation (Yarowsky, 1992; Yarowsky, 1997; Mihalcea and Moldovan, 1999; Gorman et al., 2018). These methods require large corpora that are manually tagged with correct meanings or pronunciations. The corpora are used to construct models for disambiguation and the disambiguation accuracy is strongly affected by the size of the corpora. However, manual tagging is expensive and causes serious impediments in the application of

32nd Pacific Asia Conference on Language, Information and Computation
Hong Kong, 1-3 December 2018

statistical methods to word sense/pronunciation disambiguation.

In addition, the tagged data should be balanced between the word sense/pronunciation categories to be disambiguated. As the disambiguation is a classification task, the data imbalance between the classification categories results in an overall low accuracy. To the best of our knowledge, there have been few pronunciation balanced data, at least, in Japanese, however other balanced data exist as "ATR 503 sentence" (phonetically balanced data used to train acoustic model for automatic speech recognition and speech synthesis) (Kurematsu et al., 1990) and "Balanced Corpus of Contemporary Written Japanese (BCCWJ)" (document source balanced data) (Maekawa, 2007). Although a number of electric documents are available on the World Wide Web, they do not contain pronunciation/sense tags.

A simple scaling up of the corpus is inefficient in terms of gathering examples necessary for model building, owing to the difficulty in collecting examples of sentences and readings. Word reading disambiguation requires examples of sentences that include the word whose reading is disambiguated. Zipf's Law states that few words occur frequently and many words are infrequent; therefore, if the word to be disambiguated is a rare or an infrequent word, sufficient collection is difficult. Furthermore, to construct a dataset for reading disambiguation, it is necessary to collect examples of sentences for each reading of the word in sufficient and well-balanced quantities. Some readings are special in some contexts and it is very difficult to collect the words with the readings in sufficient quantities. Thus, a data collection method focusing on a specific word is required.

Hence, this paper presents a method for acquiring pronunciation tagged data for Japanese word reading disambiguation, which is an extension of the method proposed by Mihalcea et al. (1999). The focus is on pronunciation tags instead of sense tags as investigated by Mihalcea et al. (1999). Given a word with ambiguity in pronunciation (a heteronym), in this method, the ambiguous word is first replaced by a word whose pronunciation is unique (that is, a word that has only one pronunciation (not ambiguous in terms of pronunciation)) and the same meaning as that of the original word in its context. Then, the replaced word is used as a query word to search for sentences that include the replaced words. Finally, the query word is replaced by the original word (the word with ambiguity in pronunciation) in the sentences and pronunciation tags are given to the sentences. In this way, the proposed method constructs a dataset in which words with ambiguity in pronunciation are paired with their correct pronunciations for word reading disambiguation. It is shown by experiments that the proposed method collects data efficiently and the collected data is numerically balanced among all the readings of the heteronyms.

2 Dataset construction method

The method proposed in this study enables the semi-automatic acquisition of sentences as possible examples in which a particular pronunciation of a word might occur and the word will be pronunciation tagged in all these examples. The acquisition of a pronunciation tagged dataset for a particular word, involves the following four main steps:

Step 1 Word replacement step.
A word with ambiguity in pronunciation (W_{ap}) is replaced by a word with unique pronunciation (W_{up}) whose sense is the same as that of W_{ap} and whose pronunciation is unique, that is, whose pronunciation is not the same as that of any other words.

Step 2 Search for example of a sentence step.
W_{up} is used as a query word to search for the sentences including W_{up}.

Step 3 Original word replacement and tagging step.
W_{up} is replaced by W_{ap} in the sentences retrieved in step 2. Then, the pronunciation of W_{ap} in each sentence is tagged to one of the pronunciations of W_{ap} that corresponds to W_{up} used in step 2.

Step 4 Sentence confirmation step.
Human evaluators confirm all the retrieved sentences according to the following two criteria:

1. whether the usage of the replaced W_{ap} is appropriate or not
2. whether the tagged pronunciation of W_{ap} is correct or not

in the context of the retrieved sentence.

Table 1: Example of sentence acquisition from step 1 to 4: A word with ambiguity in pronunciation (W_{ap}) "今日" has two readings /kon nichi/ (recently) and /kyoo/ (today). A retrieved example of a sentence for /kon nichi/ obtained with "最近" (recently) as a word with unique pronunciation (W_{up}) is shown in the upper half and a retrieved example of a sentence for /kyoo/ obtained with "本日" (today) as a W_{up} is shown in the lower half. The numbers at the extreme left are the line numbers. Underlined words in line 2/9 are replaced by those in line 3/10. An English translation of the Japanese sentences in lines 2 and 3 is given in line 4 and that of the Japanese sentences in line 9 and 10 is given in line 11 and the underlined English words correspond to W_{ap} in Japanese.

#	step	processing results
1	step1	W_{ap}="今日" (/kon nichi/) → W_{up}="最近" meaning "recently"
2	step2	昨日は<u>最近</u>の寒の戻りの特異日でしたが · · ·
3	step3	昨日は<u>今日</u>の寒の戻りの特異日でしたが · · ·
4		= Yesterday was a unique day of <u>recent</u> cold weather · · ·
5		pronunciation tag is set to /kon nichi/
6	step4	Judged "Not appropriate"
7	*final result*	discard the sentence obtained in step 3
8	step1	W_{ap}="今日" (/kyoo/) → W_{up}="本日" meaning "today"
9	step2	私も<u>本日</u>気になったので、本屋をのぞいてみても · · ·
10	step3	私も<u>今日</u>気になったので、本屋をのぞいてみても · · · · · ·
11		= I was also interested <u>today</u>, I looked into the bookstore · · ·
12		pronunciation tag is set to /kyoo/
13	step4	Judged "Appropriate"
14	*final result*	pronunciation /kyoo/ is tagged to the sentence obtained in step 3

The four steps outlined above are applied for each pronunciation of W_{ap}. The details of the steps are described in the following subsections and Table 1 shows examples of the processing results at each step.

2.1 Word replacement step (Step 1)

This step replaces the word with ambiguity in pronunciation (W_{ap}) by a word with unique pronunciation (W_{up}). Wordnet is used (Miller et al., 1994) to find the word (W_{up}) whose sense is the same as W_{ap} and whose pronunciation is unique. If the appropriate word is not found in WordNet, W_{up} is manually specified.

Examples are given below using a Japanese word "今日" as a word with ambiguity in pronunciation (W_{ap}). Hence we use the Japanese WordNet (Isahara et al., 2008). The Japanese word "今日" is a heteronym, i.e., the word whose pronunciation is /kon nichi/ means "recently," whereas the word whose pronunciation is /kyoo/ means "today." As the word with both pronunciations functions as an adverb in sentences, it is necessary to disambiguate the reading based on information other than the part of speech. To obtain examples of sentences for both the pronunciations, "今日" is replaced by "最近" which means "recently" and whose pronunciation is unique (/saikin/), and also by "本日" which means "today" and whose pronunciation is also unique (/hon jitsu/).

These replaced words, "最近" and "本日" can be found in the Japanese WordNet (Isahara et al., 2008). In the WordNet, each word belongs to a synonym set, "synset," which includes several synonyms. Semantically different words belong to different synsets. The synsets have a hierarchical structure, e.g., "今日" belongs to an upper layer (synset id=15119536-n) and a lower layer (15156001-n and 15262921-n); "今日" meaning "today" belongs to a lower synset (15156001-n) different from the lower synset of "今日" meaning "recently" (15262921-n). By referring to each lower layer of the synset, words with unique pronunciations (W_{up}) can be automatically selected as query words. These replacements are illustrated in line 1 and 8 in Table 1. If an appropriate word is not found in the WordNet, W_{up} is manually specified with due attention paid to both pronunciation uniqueness and semantic sameness.

2.2 Search for example of sentence step (Step 2)

This step uses the replaced word with unique pronunciation (W_{up}) as a query word to search for examples of sentences that include the query word from the World Wide Web or already collected electrical documents.

The word "今日" used in step 1 (as a W_{ap}) is replaced by either "最近 (recently)" or "本日 (today)" as W_{up} in step 1 as illustrated in line 1 and 8 in Table 1. W_{up} is used as a query word to search for sentences that include W_{up}. The search results of examples of sentences are shown in lines 2 and 9 and the query words are underlined (最近 in line 2 and 本日 in line 9). The English translations of both are given in lines 4 and 11.

2.3 Original word replacement and tagging step (Step 3)

This step replaces W_{up} by the original W_{ap} in the sentences obtained in the search. In addition, the pronunciation of W_{ap} in each sentence is tagged to one of the pronunciations of W_{ap} that corresponds to W_{up} used in step 2.

In the example in Table 1, the original W_{ap} was "今日". The word underlined in line 2 (最近) is replaced by the original word underlined in line 3 (今日) and the underlined word in line 9 (本日) by that in line 10 (今日). These sentences are also tagged with pronunciations, e.g., /kon nichi/ applies to the sentence in line 3 as well as in line 5, and /kyoo/ applies to the sentence in line 10 as well as in line 12.

2.4 Sentence confirmation step (Step 4)

In this step, the sentences that are obtained are confirmed manually. Each sentence is confirmed in the context of the retrieved sentence, in accordance with the following two criteria: 1) whether the usage of the replaced W_{ap} is appropriate or not, and 2) whether the tagged pronunciation of W_{ap} is correct or not.

As W_{ap} is not included in the sentences collected in the search, artificial sentences are constructed by the replacement of W_{up} by W_{ap}. A human evaluator judges whether the usage of W_{ap} in the retrieved sentence is appropriate or not (the first criterion). The sentence which is found appropriate in the first evaluation is then judged in terms of the correctness of the tagged pronunciation in the sentence (the second criterion).

Examples of the confirmation results are shown in line 6 ("Not appropriate" for the sentence in line 3) and in line 13 ("Appropriate" for the sentence in line 10) of Table 1. Finally, the sentence that is judged to be appropriate is saved as sentence data for word reading disambiguation and the sentence judged to be inappropriate is discarded.

3 Experiment

The acquisition efficiency of the proposed method was evaluated and compared to that of a conventional dataset construction procedure.

3.1 Conditions

3.1.1 Target words

UniDic (Den, 2009), an electric dictionary, which is publicly available, was used, from which word pairs of the same word form and part-of-speech, but having different pronunciation were extracted. In this way, 552 pairs of 2,128 words were obtained, which were classified into four classes as shown in Table 2.

A heteronym is important for speech synthesis and corresponds to category 1 and 3 in Table 2, whereas meaning in category 1 is an exclusive relationship, as one reading does not match another, however, in category 3, some meanings are the same as others while some are not. As shown in category 3 in Table 2, "縁" has at least two readings of /en/ and /fuchi/. In the case that "縁" is used with /en/ meaning *edge*, both the readings of /en/ and /fuchi/ might be allowed, but in the case that "縁" is used with /en/ meaning *emotional ties*, it should be read as /en/. In this study, the focus is on heteronyms having exclusive differences in meaning as shown in Table 2 and 43 pairs of 88 words in category 1 of the Table 2 were used in the following experiments. Although more than one word with a unique pronunciation (W_{up}) can be used for each of the 88 words to collect a variety of sentences, only one W_{up} has been used for each word in the following experiments.

Table 2: Classification of words whose word forms are the same but whose pronunciations differ

#	meaning	difference in	example
1)	different	meaing	/kyoo/ (today) or /kon nichi/ (recently) of 今日
2)	same	voiced or unvoiced consonant	/kaisya/ or /gaisya/ of 会社 (company)
3)	same/different	Chinese or Japanese reading	/en/ or /fuchi/ of 縁 (edge)
	same	Chinese or Japanese reading	/kafuku/ or /shita bara/ of 下腹 (inferior abdomen)
4)	same	colloquial or literary	/iku/ or /yuku/ of 行く (to go)

3.1.2 Sentence source and amount of collection

Although crawlers and web search engines can be used in these experiments to collect sentences, a publicly available web text corpus was utilized for reproducibility of the experiments and in this study the "Text Archive Japanese Web Corpus 2010[1]" was used. This corpus consists of sentences randomly collected by using the search engine of Web document space, and includes collected web documents for each input word of the ipadic-2.7.0 (Asahara and Matsumoto, 2003) , which is commonly used in Japanese, consisting of 400,000 different words. The amount of text in it is 396 GB (billions of sentences), which is considered to be sufficiently large as a subset of web documents to confirm the efficacy of the proposal made in this study.

We collected up to 100 sentences for each pronunciation from this source.

3.1.3 Human evaluators

Each collected sentence was confirmed by two human evaluators in accordance with the aforementioned criteria and counseling was prohibited between the two evaluators.

3.1.4 Evaluation measure

An adoption rate was defined as given below, to evaluate the acquisition efficiency of the proposed method:

$$adoption\ rate[\%] = \frac{passed}{all} \times 100.0 \qquad (1)$$

where *all* denotes the total number of sentences collected by the method used in this study and *passed* denotes the number of sentences satisfying the two criteria stated in step 4.

[1]http://s-yata.jp/corpus/nwc2010/

3.2 Results

3.2.1 Acquisition efficiency

The acquisition efficiency was first determined. We focused on the words for which our method exactly collected 100 sentences. Thus, adoption rates were calculated under the condition that *passed* is set to 100 in equation (1). The second column from the extreme right in Table 3 shows a summary of the adoption rates calculated in this study. The number of words for which 100 sentences were exactly collected was 20 and the number of words for pronunciation was 40 as shown in Table 3 and the average adoption rate among the readings was 74.0%.

3.2.2 Comparison

The conventional data construction procedure must be reproduced, with the collected sentences, including words with ambiguity in pronunciation (W_{ap}) and the correct pronunciation of the words manually tagged in the sentences, for an ideal comparison. However, this kind of manual tagging of pronunciation is expensive.

Therefore, a similar procedure was simulated by collecting sentences, that included words with ambiguity in pronunciation (W_{ap}), from the large existing database containing the correct pronunciation manually tagged and classifying the pronunciations in the sentences based on the tagged correct pronunciation. The "Balanced Corpus of Contemporary Written Japanese (BCCWJ)" (Maekawa, 2007) was used, as it consists of 60,000 sentences and its core portion includes correct readings. We randomly collected up to 100 sentences for each pronunciation of the words with ambiguity in pronunciation and calculated the adoption rates for each reading by using equation (1).

The extreme right column of Table 3 shows the

Table 3: Adoption rate comparison: The word with ambiguity in pronunciation, W_{ap}, in the extreme left column has several readings given in the second column and the corresponding meanings or usage given in the third column from the extreme left column, respectively, whereas the two columns on the right include the adoption rates.

W_{ap}	**reading**	meaning in English or usage	**adoption rate [%] by our method**	**by simulated conventional procedure**
半月	/han tsuki/	half a month	100	39
	/han getsu/	half moon	99	3
金	/kin/	gold	81	100
	/kane/	money	97	42
今日	/kyoo/	today	99	100
	/kon nichi/	recently	87	32
質	/shichi/	gage	83	100
	/shitsu/	quality	97	6
便	/bin/	service	95	19
	/ben/	usability	98	2
造作	/zoosa/	trouble	98	17
	/zoosaku/	feature	91	9
寒気	/samuke/	chill	100	3
	/kanki/	cold air	100	6
訳	/wake/	reason	99	3
	/yaku/	translation	91	2
大事	/daiji/	important affair	97	1
	/oogoto/	serious	67	100
目下	/meshita/	junior	97	4
	/mokka/	nonce	72	2
品	/shina/	goods	94	1
	/hin/	elegance	77	100
後	/ushiro/	backward	92	4
	/ato/	residue	53	100
空	/kara/	empties	86	1
	/sora/	sky	57	45
入り	/hairi/	start,	64	57
	/iri/	revenue	52	29
人気	/hitoke/	a sign of life	58	46
	/ninki/	popularity	55	2
方	/hoo/	direction/choice	77	100
	/kata/	person	98	100
実物	/jitsubutu/	real thing	73	3
	/mimono/	kerneled thing	0	1
避け	/sake/	keep off	57	100
	/yoke/	dodge	16	1
人	/jin/	a postfix showing that persons have some characteristics as 国際人 (internationally minded person)	36	100
	/nin/	a postfix showing person's work as 弁護人 (counsel)	3	100
床	/toko/	bed/alcove	28	3
	/yuka/	floor	35	33
average adoption rate [%]			74.0	37.9
average adoption rate difference between readings [%]			19.5	41.4

adoption rates of the sentences obtained by the simulated conventional dataset construction procedure with BCCWJ, and the average of the adoption rates was calculated.

The adoption rate for the method used in this study was 74.0%, which was 36.1 points higher than that obtained for the simulated conventional procedure (37.9%). This suggests that the conventional dataset construction procedures by classification after collection result in lower adoption rates and may not be efficient.

As explained above, the number of sentences should be balanced among the pronunciation categories. The calculated difference in the average adoption rate between the pronunciations of each word is shown at the bottom of Table 3. As the difference shown by the proposed method and simulated conventional procedure is 19.5% and 41.4%, the former produced a balanced dataset for pronunciation, with higher acquisition efficiency.

3.3 Discussion

3.3.1 Total adoption efficiency

The adoption rates for all the words with ambiguity in pronunciation were also calculated by including the words for which fewer than 100 sentences were collected. The average adoption rate of the present method was 58.2% and that of the simulated conventional procedure was 24.7%. In addition, there was an average difference of 30.8% between the adoption rates for pronunciations by the present method and 25.3% by the simulated procedures. The smaller difference in the adoption rate of

the simulated procedures may be due to the smaller number of collected sentences and is not considered to be critical. These results suggest that the proposed method might be more efficient in dataset construction than the conventional one.

3.3.2 Query word selection

The selection of words with unique pronunciation might be the key for the success of the proposed method. This is because higher adoption rates were obtained when the appropriate word with unique pronunciation was selected. The appropriate word with unique pronunciation is the word whose pronunciation is different from that of any of the other words. The examples of successful selection were found in the case of "今日", "寒気 (/kanki/ (cold air) or /samuke/ (chill))" and "半月 (/han getsu/ (half moon) or /han tsuki/ (half a month))."

However, a failed example was found in a short word such as "空 /sora/ (sky)," "人 (/jin/ or /nin/)," and "床 (/toko/ or /yuka/)." For example, "人" is a postfix showing either the job or characteristics of a person. "料理人 (cooking person)" and "弁護人 (counsel)" signifies the work of a person, a cooking person and a person putting forth a defense, respectively, and "人" in both these words is read as /nin/, while "国際人 (international people)" and "知識人 (a man of knowledge)" signifies characterized persons, an internationally minded person and an intellectual, respectively and "人" in both these words is read as /jin/. This disambiguation might be possible by focusing on the attribute of the previous word as "料理 (cooking)," "弁護 (pleading)," "知識 (knowledge)," and "国際 (international)." As the reading /nin/ might follow a verbal noun such as "料理" and "弁護" a statistical language model using adjacent words and their attributes (verbal noun or not) may be favorable for disambiguation. Although longer expressions such as "料理する人 (person who is cooking)" and "知識のある人 (people who have knowledge)" mean the same as the shorter expressions of "料理人" and "知識人" , they may not be as commonly found in natural documents as compared to the shorter expressions.

3.3.3 Replacing the original word

Replacing a short word by a longer one may be effective when an appropriate and frequently used word is found. For example, 空 (/kara/ (empties) or /sora/ (sky)) can be replaced by "空っぽ" /karaqpo/ (empties) or "お空" /o sora/ (sky, a courteous word for sky), 床 (/toko/ (bed/alcove) or /yuka/ (floor)) to "寝床" /ne doko/ (bed) or "フロアー" /furoaa/ (floor). However, the longer word sometimes matches beyond the correct word boundaries. Thus, care must be taken not to make a mistake while replacing a short word. One meaning of 人気 is "a sign of life" for the reading /hitoke /, thus, it can be replaced by "人の気配 (people's signs)" as a query word in step 1, but the portion "気配" of the replaced word has two readings as /kehai/ (a sign of life) and /kikuba(ri)/ (caring nature), hence the method retrieves the sentence including "支配人の気配りで (caring nature of manager)" (the underlined part is the part that matches the replaced word(人の気配)) and fails when the query is replaced by the original word as "支配人気で (the English translation is omitted because the word sequence is unnatural in Japanese.)" (the underlined part is the replaced original word). A replacement for longer words is left for our work in future.

3.3.4 Use of crowdsourcing

The use of crowdsourcing for database construction is discussed here. Persons may be asked to construct sentences that include words with ambiguity in pronunciation and to tag the correct pronunciations. However, limits are anticipated for the number and varieties of sentences that one person can write. At the moment, it is considered that easier tasks might be appropriate for the use of crowdsourcing than writing pronunciation tagged sentences. Hence, we believe that the proposed method in this study has some significance for pronunciation tagged corpus construction.

3.3.5 Usefulness of acquired data

Word reading disambiguation for all the words with pronunciation ambiguity must be performed to confirm the usefulness of the acquired data. However, a preliminary word reading disambiguation experiment was conducted where "方" with reading ambiguity of /hoo/ and /kata/ was used. For each reading 274 sentences were taken from the BCCWJ database and divided into 5 sets. Four sets were used as training data and the remaining set was used as

test data for 5-fold cross validation. Classification features used by Gorman(2018) were followed and unigram and bigram of the word surface form and part-of-speech were used before and after the target word to be disambiguated. This experiment used the conditional random field (Lafferty et al., 2001) as a simple classifier and an average correct rate of 76.6% (with a variance of 4.6) was obtained.

The proposed method was used and 77 sentences were collected for each reading of the target word "方" and 77 sentences were added to each training set for cross validation. This resulted in an average correct rate of 78.3% (with a variance of 10.1). Statistical tests showed that the difference in the average correct rate was statistically significant ($p<0.01$). These preliminary results appear to be promising and the proposed method can be considered to contribute favorably to useful data acquisition for word reading disambiguation. Validation on other remaining words is a feature work.

4 Conclusion

A data construction method was proposed for word reading disambiguation. We expect that similar words appear in similar sentences and thus, in the proposed method, a word with ambiguity in pronunciation (heteronym) is replaced by a word whose pronunciation is unique and the meaning is the same as that of the ambiguous word. The unique word is then used as a query word for searching for sentences that include them, which is then replaced by the original ambiguous word to construct pronunciation tagged corpora. It was confirmed by experiments that the proposed method was more efficient than the conventional dataset construction procedures and was numerically balanced among the readings of heteronyms and the collected data provided a statistically significant improvement in preliminary disambiguation experiment.

References

Keith Hall and Richard Sproat. 2013. *Russian stress prediction using maximum entropy ranking*. Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing, 879–883.

Richard Sproat and Keith Hall. 2014. *Applications of maximum entropy rankers to problems in spoken language processing*. Proc. of INTERSPEECH, 761–764.

Daniela Braga, Luis Coelho, and Fernando Gil V. Resende Jr.. 2007. *Homograph ambiguity resolution in front-end design for Portuguese TTS systems*, Proc. of INTERSPEECH, 1761–1764.

Kyle Gorman, Gleb Mazovetskiy, and Vitaly Nikolaev. 2018. *Improving homograph disambiguation with supervised machine learning*, Proc. of LREC, 1349–1352.

David Yarowsky. 1997. *Homograph disambiguation in text-to-speech synthesis*. Progress in speech synthesis, Springer, New York, 157–172.

David Yarowsky. 1992. *Word-sense disambiguation using statistical models of Roget's categories trained on large corpora*. Proc. of COLING, 454–460.

Rada Mihalcea and Dan I. Moldovan. 1999. *An automatic method for generating sense tagged corpora*. Proc. of AAAI, 461–466.

Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano. 1990. *ATR Japanese speech database as a tool of speech recognition and synthesis*. Speech Commnication, 9, 357–363.

Kikuo Maekawa. 2007. *Kotonoha and BCCWJ: development of a balanced corpus of contemporary written Japanese*. Proc. of the First International Conference on Korean Language, Literature, and Culture, 158–177.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2009. *Development of Japanese WordNet*. Proc. of LREC, 2420–2423.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. *Using a semantic concordance for sense identification*. Proceedings of the ARPA Human Language Technology Workshop, 240–243.

Yasuharu Den. 2009. *A multi-purpose electronic dictionary for morphological analyzers*. Journal of Japanese Society for Artificial Intelligence, Vol. 24, No.5, 640–646 [in Japanese].

John Lafferty, Andrew McCallum, Fernando C.N. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Probabilistic models for segmenting and labeling sequence data, 282–289.

Masayuki Asahara and Yuji Matsumoto 2003. *ipadic version 2.7.0 User's Manual*. Nara Institute of Science and Technology.