Orthographic variation problems and the Japanese Wordnet

Takayuki KURIBAYASHI
Division of Linguistics and Multilingual Studies
Nanyang Technological University

At the setout

- What is "orthographic variation"?
  - Words can be written in more than one form
  - Orthographic variants have the same meaning and reading in common
- Not so many patterns In English

  e.g.

  center / centre

  color / colour

Source of orthographic variation problems in Japanese

The 3 scripts in Japanese

- Kanji ("漢字", Chinese character)
  - Ideogram
  - Sometimes has different shapes and combined in a string

    e.g. for *gakkou* (school)

    {
    New letter shape "学校"

    Old letter shape "學校"
- Kana
  - Phonogram
  - 2 types
    - Katakana "カタカナ"
    - Hiragana "ひらがな"

Choice of the script(s)

- In modern Japanese, a word string usually consists of a single script or kanji + hiragana
- A choice of scripts depends on the writer and type of document
  - [”犬”, ”イヌ”, ”いぬ”] for ”dog”
  - In informal documents such as novels and blogs, it more depends on the writer

Kanji + hiragana string

- Kanjis often need okurigana (送り仮名, accompany letters)
  - In the first place, Japanese readings can not fit the kanji's original readings
  - Most kanjis have more than one meaning
    - Okurigana is needed to reduce the ambiguity

Examples of okurigana

- "重" oroginal readings: *juu*, *chou*

"重"	*e*, *juu*	*(numeral classifier)*
"重い"	*omo-i*	heavy
"重さ"	*omo-sa*	weight
"重ねる"	*kasa-neru*	pile
"重なる"	*kasa-naru*	overlap
"重ねて"	*kasa-nete*	again

Okurigana rules

- The Japanese government has issued a guideline for okurigana
  - But only reveals in newspapers, official documents, legal sentences, and so on
- No strict rule for usage in other kinds of writings
  - Conjugation part can not be omitted
    - "重い", ”重ねる”, ”重なる"
  - Not recommended to omit if the disambiguation is obstructed
    - Which does ?"重る" means?

Sources of orthgraphic variation (review)

- Freely decided which script to use
  - Scripts : kanji, katakana and hiragana
  - Kanjis often need okurigana
    - How many okuriganas to use is relatively free, too
  - Choices are depend on the type of the document and/or the writer’s liking

Other examples of variation

- ”おそろしい(*osoroshii*), terrible”

  ”恐ろしい”, ”恐しい”, ”オソロシイ”, ”おそろしい”

- ”ひふ(*hifu*), skin”

  ”皮膚”, ”皮フ”, ”皮ふ”, ”ヒフ”, ”ひふ”

- ”まぜあわせる(*mazeawaseru*), mix”   consists of ”まぜる” & ”あわせる” = 32 variants
  - ”まぜる(*mazeru*), mix”

    ”混ぜる”, ”交ぜる”, ”雑ぜる”, ”混る”, ”交る”, ”雑る”, ”マゼル”, ”まぜる”

  - “あわせる(*awaseru*), combine”

    ”合わせる”, ”合せる”, ”アワセル”, ”あわせる”

Actual problems

Actual problems

1. Japanese Wordnet (JWN) 1.1 does not cover all the variants
   - Affect the coverages when annotating corpora
2. A variant sometimes appears in a synset, but misses in other synsets

   e.g. “吸い込む”

3. Are the numbers of synonyms and senses (synonym-synset pair) reasonable?

   we counted "吸い込む", "吸込む" separately

1.Strings not covered when annotating

- In a newspaper corpus (Kyoto University Text Corpus)
  - "防空ごう(*boukuu-gou)*, bombproof",
    we have "防空壕" in 02868638-n
  - "あやうい(ayaui), dangerous",
    we have "危うい" in 02058794-a
- In a novel
- In a old Japanese novel
  - Some Meiji era novelists preferred ”恐しい” than ”恐ろしい”?

Actual problems

1. Japanese Wordnet (JWN) 1.1 does not cover all the variants

   Affect the coverages when annotating corpora

2. A variant sometimes appears in a synset, but misses in other synsets

   e.g. ”吸い込む” appears in 6 synsets

   ”吸込む” appears in 5 synsets

3. Are the numbers of synonyms and senses (synonym-synset pair) reasonable?

   we counted "吸い込む", "吸込む" separately

Actual problems

1. Japanese Wordnet (JWN) 1.1 does not cover all the variants

    Affect the coverages when annotating corpora

2. A variant sometimes appears in a synset, but misses in other synsets

    e.g. “吸い込む”

3. Are the numbers of synonyms and senses (synonym-synset pair) reasonable?

    e.g. we counted "吸い込む", "吸込む" separately

To solve the problem

Our method

1. Create variant sets with help from open-licenced dictionaries
2. Apply the variant sets to JWN 1.1 synonyms
3. Hand check

➡ adding & grouping variants

Dictionaries

- 3 dictionaries
  - JUMANdic by Kyoto University
    - For their morphological analysis system JUMAN
    - Entries can be grouped by canonical form & reading
  - JMdict managed by EDRGD
    - Entries can be grouped by meaning & reading
  - IPAdic by NAIST
    - We hired merely to give reading the synonyms not in JUMANdic nor JMdict

Merging 2 dictionaries

- Merge the JUMANdic entries and JMdict entries that can be identified as the same word or its variants
  - e.g. ”荒らす(*arasu*), desolate”

{ JUMANdic: [荒らす, あらす]
  JMdict : [荒らす, 荒す, あらす]

➡ merged : [荒らす, 荒す, あらす]

Giving reading

- Give the each merged set a katakana string as reading
- By converting the hiragana string in JMdict

  - e.g. ”あらす”

    ”あらす” → ”アラス”

Why do we need kana strings?

- Kana is made available as phonogram in Japanese, therefore adding reading information is equal to adding kana strings
- On top of that, the difference of reading can contribute Word Sense Disambiguation (WSD) in some cases
    - e.g. “面” can be read as:
        - a) “ツラ(*tsura*)”, ”オモテ(*omote*)”, ”メン(*men*)”
        - b) “メン(*men*)”

Giving reading (cont’d)

- If a synonym is not in JUMANdic nor JMdict, do morphological analysis and give them the readings with IPAdic
  - e.g. ”情報機関(*jouhoukikan*), intelligent agent”

IPAdic: [情報, じょうほう] + [機関, きかん]

⬇

[情報機関, ジョウホウキカン, じょうほうきかん]

In case readings are not found

- Give the synonyms a tag that means ”its reading is unknown”

  e.g. “吹弾 (*suidan*), play in 01725051-v)”

  [吹弾, YOMI, YOMI]

Deciding display form

- Decide a display form for each variant set
  - We do not say "standard form" since no one can decide undisputed ones
  - Merely in order to create a key for each set
    - Show only one form when searching JWN
    - Use for sentence generating

表示表記決定優先度

1. Has the highest frequency --- N/A as of now
2. Agrees with JUMANdic's canonical form
3. Consists of more chinese characters
4. Consists of more new letter shape ones
5. Is longer if 1 ~ 4 can not settle

Create the key

- To make a variant set's ID, give each display form one digit
  - This is to deal with variant sets which have the same display form like ”面"
  - e.g. ”荒らす”

[荒らす 0, アラス, 荒す, あらす]

key reading

==> Hand check all variant sets (done)

Apply variant sets

- Apply the hand-checked variant sets to JWN 1.1 synonyms
  - when a synonym is in the variant sets, we apply the sets
    - e.g. ”面” appears in 6 variant sets and each JWN synset which has ”面” are applied 6 sets
- Hand check again to remove variant sets which are applied incorrectly

  e.g. ”面” in 03724870-n (”mask”)

{ ◎ 面, メン, めん read as ”*men*”
  ✖ 面, ツラ, 頬, つら read as ”*tsura*”

Status of the JWN (as of Jan 2016)

- 91,961 unique words → 83,174 variant sets
  213,986 unique strings

- 158,074 senses (synset-synonym pairs) →
  148,005 synset-variant set pairs
  449,240 synset-string pairs

(the numbers include error correction)

Examples

- [みみずく 0 (ミミズク, 木兎, 角鴟, 木菟)]
- 02765464-v (”absorb”, ”take in”)

  JWN 1.1 : 呑み込む, 呑みこむ, 呑込む, 吸引, 吸い込む,吸収

⬇

吸い込む	スイコム, 吸込む, 吸いこむ, すいこむ
吸収	キュウシュウ, きゅうしゅう
吸引	キュウイン, きゅういん
飲み込む	ノミコム, 飲込む, 呑み込む, 呑込む, のみ込む, のみこむ
呑みこむ	ノミコム, のみこむ

results of applying

Coverage (as of 2012)

	Total words	Content words	Covered content words	Coverage
Dancing Men	13,483	4,752	3,874	81.5%
			4,332	91.2%
Speckled Band	13,896	4,848	4,097	84.5%
			4,501	92.8%
Cathedral & Bazaar	18,067	7,509	5,858	78.0%
			6,618	88.1%
Kyoto Corpus (articles)	24,615	11,939	9,385	78.6%
			9,766	81.8%
Kyoto Corpus (editorial)	27,906	13,300	10,958	82.4%
			11,542	86.8%

results of applying

Problems and future work

Increased ambiguity

1. The hand checking takes time
    - The data before checking contained many errors which come from ambiguity since we considered improving the coverage first
    - Especially kana strings increase ambiguity

      e.g. Each ”タイ (*tai*)” in JWN 1.1 is applied
      10 variant sets before checking

Rare forms

2. A variant set contains rare forms in some cases and increase ambiguity

- Rare ones should be removed or suppressed to appear by using frequency data in the future

  e.g. ”頬” in the variant set ”面 (*tsura*)”

Need to further merge

3. Not all the variants are merged into each variant set

- Target : strings which are not in JUMANdic nor JMdic
- If the variant sets which appear in the same synset and have the same reading in common should be merged (such as ”呑みこむ” in 02765464-v, pp29)

Reading (kana strings) information is important also in this respect

Relationship with OMW

4. This attempt has proceeded independently of our Open Multilingual Wordnet
   - Error correction in both side independently
   - How to merge the data?

Conclusion

- We need to handle orthographic variants
- Without them, our coverage is poor
- We need to group variants
- We do this by
  - Find dictionar(ies) in which orthographic variants are grouped
  - Connect the dictionar(ies) to your Wordnet by reading information
  - Checking them