Automatic Acquisition of Morphological Modification Rules to Look up Entries in Lexicons

Masaru Morita (110119)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 2003

Keywords: Morphological Analyzer, Lexicon, Morphological Modification Rules, Strings of Morpheme, Morpheme Boundary.

In natural language processing, semantic information (e.g. semantic classes or word sense descriptions) of morphemes in sentences is often looked up in lexicons (e.g. thesauri or dictionaries). Furthermore, sentences are often segmented into morphemes using morphological analyzers as pre-processing of looking up entries in lexicons. However, lexical entries, such as semantic classes in thesauri or word sense descriptions in dictionaries, may not be retrieved because of the disagreement between strings(written in text) of morphemes outputed by morphological analyzers and strings(headwords) of lexical entries, and the disagreement of morpheme boundaries in morphological analyzers and lexicons. Such disagreement of strings and boundaries of morphemes should be modified in order to retrieve lexical entries from lexicons efficiently. However, there are several morphological analyzers and lexicons which are available to researchers, it is required too much labor and time to construct modification rules of strings or boundaries of morphemes for all combinations of morphological analyzers and lexicons. This paper aims at acquiring modification rules automatically, which modify strings or boundaries of morphemes outputed by a morphological analyzer into ones in a lexicon, by checking morpheme list of a dictionary of morphological analyzer and a lexicon.

Two types of modification rules are acquired in this research.

- Rules modifying strings of morphemes

  They modify strings of morphemes outputed by a morphological analyzer into strings in a lexicon. "輪なげ (*wanage*, quoits) → "輪投げ (*wanage*, quoits)" is an example of the rules. This rule modifies the string "輪なげ" outputed by a morphological analyzer into the string "輪投げ" which is the headword in a lexicon. The rules are referred to as "one-to-one rules".

- Rules modifying boundaries of morphemes

  First, the rules which divide a morpheme outputed by a morphological analyzer into two or more morphemes in a lexicon, referred to as "one-to-many rules". "大量消費 (*tairyoshohi*, mass consumption)" → "大量 (*tairyo*, mass)" + "消費 (*shohi*,

consumption)" is an example of the rules. This rule divides the morpheme "大量消費" outputed by a morphological analyzer into two morphemes "大量" and "消費" in a lexicon.

Next, the rules which concatenate several morphemes outputed by a morphological analyzer into one morpheme in a lexicon, referred to as "many-to-one rules". "経済 (*keizai*, economy)" + "成長 (*seityo*, growth)" → "経済成長 (*keizaiseityo*, economical growth)" is an example of the rules. This rule concatenates two morphemes "経済" and "成長" outputed by a morphological analyzer into a morpheme "経済成長" in a lexicon.

The method to acquire one-to-one rules is as follows. First, suppose the set of morphemes in a dictionary of a morphological analyzer to be $M = \{(h_m, y_m, p_m)\}$, and the set of morphemes in a lexicon to be $D = \{(h_d, y_d, p_d)\}$. Each morpheme in M and D is represented by a triple of string h, pronunciation y and part of speech (POS hereafter) p. As sets of POSs of a morphological analyzer and a lexicon are different in general, each POS in a morphological analyzer and a lexicon is converted to one of the coarse set of common POSs, such as "noun", "verb" etc. Next, find the pair of morphemes (h_m, y_m, p_m) and (h_d, y_d, p_d) where $y_m = y_d$, $p_m = p_d$, h_m and h_d are matched, then acquire an one-to-one rule $(h_m, y_m, p_m) \rightarrow (h_d, y_d, p_d)$. The precious definition of "h_m and h_d are matched" in above is that two strings are matched by DP matching under the following condition: (1) the same character is matched each other, (2) one kanji character is matched to any string of hiragana.

The method to acquire one-to-many rules are as follows. First, proper nouns are eliminated from M, because it is meaningless to divide proper nouns. Next, find the set of morphemes (h_{d1}, y_{d1}, p_{d1}), $\cdots$, (h_{dn}, y_{dn}, p_{dn}) for each morpheme (h_m, y_m, p_m) in M satisfying the following 4 conditions, then acquire an one-to-many rule $(h_m, y_m, p_m) \rightarrow (h_{d1}, y_{d1}, p_{d1}) + \cdots + (h_{dn}, y_{dn}, p_{dn})$. (1) $h_m = h_{d1} \oplus \cdots \oplus h_{dn}$ ($\oplus$ indicates concatenation of strings), (2) $p_m = p_{d1} = \cdots = p_{dn}$, (3) h_m doesn't contain any hiragana or symbols, (4) the length of one of h_{di} should be more than two. The method to acquire many-to-one rule is almost same as one-to-many rules, except for exchanging M and D. The only difference is that any proper nouns doesn't eliminated from M.

Modification rules are acquired for 8 combination of two morphological analyzers (JUMAN and Chasen) and four lexicons (Iwanami Kokugo Jiten, Bunrui Goi Hyo, Nihongo Goi Taikei and EDR Japanese dictionary). 11,000 $\sim$ 300,000 one-to-one rules are acquired, where 97 $\sim$ 99% rules are appropriate as one-to-one rules. On the other hand, 100 $\sim$ 21,000 one-to-many or many-to-one rules are acquired, which is much less than one-to-one rules. 80% $\sim$ 90% of one-to-many rules are correct when 50 rules are selected at random and checked by hand. Many-to-one rules concatenate several morphemes outputed by a morphological analyzer into one morpheme in a lexicon, and it is always regarded that such operations are appropriate. Thus all many-to-one rules are correct.

In order to evaluate the effectiveness of acquired modification rules, 10,000 Mainichi Shimbun newspaper articles in 1997 are morphologically analyzed, then modification rules are applied and the gain of the number of morphemes which succeed to retrieve lexical entries from a lexicon are measured. The gain is 2 $\sim$ 7% for one-to-one rules. On the other hand, there is no remarkable gain for one-to-many and many-to-one rules. This is because the number of one-to-many and many-to-one rules is too small.