PUNDA Numbears: Proposal of *Goroawase* Generating System for Japanese

Pawel Dybala　Michal Ptaszynski　Rafal Rzepka　Kenji Araki
Graduate School of Science and Technology
Hokkaido University
{paweldybala, ptaszynski, kabura, araki}@media.eng.hokudai.ac.jp

Abstract:
Goroawase are Japanese word plays basing on combining various readings of numbers and symbols. This mechanism makes them potentially possible to compute and process with tools and methods of NLP. In this paper we give a brief description of the genre, analyze its mechanism, and propose an idea of PUNDA Numbears – a *goroawase* generating system. Being developed as a part of PUNDA - our research project on pun telling conversational systems for Japanese - PUNDA Numbears should be able to generate *goroawase* (numbers) propositions towards given sets of phrases. We describe particular parts of the algorithm and give an example of how it could actually work.

Keywords: *goroawase*, word plays, humor processing, language generation

1. Introduction

Goroawase is a type of Japanese word play, in which various readings of numbers and symbols are used to form words or phrases. They are quite common in Japan, often used when a memorization of sets of numbers is required. *Goroawase* are popular in the world of advertisement, in which they can be used to make phone numbers easier to remember.

In this paper we briefly describe the genre of *goroawase* and its place among other word plays in Japanese. Then we present our research project, focused on creating a Japanese pun generator, and its relation to this work. Next, we discuss possibilities of creating a *goroawase* generating computer system, and propose our idea of such a system, along with its algorithm outline and possible applications. Finally, we conclude the work pointing out some directions for the future.

2. Background

Below we analyze the phenomenon of Japanese word plays and *goroawase* as one of its genres.

2.1 Japanese – language of homophones

Word plays or puns are present in most (if not all) of existing languages. What characterizes them is that they are more or less based on various aspects of language, using such features as homophony as the source of humorous ambiguation. In some languages, however, possibilities of creating word plays are wider than in others – one such language is Japanese, as it contains relatively high amount of homophones and homophonic phrases.

2.2 Japanese word plays

Probably the most common type of word play (and of jokes in general) in Japanese is called *dajare*. A classical example of this type is *Futon ga futtonda* (*Futon*[1] flew away), in which we can see a partial homophony between the words *futon* and *futtonda* (flew away).

Although in this paper we do not intend to define and categorize the whole class of *dajare*, we assume that *goroawase* can be seen as a particular genre of *dajare*, as their general mechanisms are quite similar. What is characteristic to *goroawase* is the fact that their subjects are mostly numbers and

[1] *futon* – jap. bedding, cover

symbols.

2.3 *Goroawase* – usability and mechanisms

This particular type of Japanese word plays is commonly used to memorize strings of numbers, such as dates or phone numbers. The main mechanism used here is based on combining different readings or parts of readings of numbers and symbols so that they form words or phrases, which, if possible, should be semantically related to the base string.

Possible readings of Japanese numbers are presented in Table 1[2].

Number	Readings
0	maru, ma, rei, re, nai, na, wa, o, zero, ze
1	ichi, i, hitotsu, hito, hi, bi, pi, fi, wan, a
2	ni, ji, futatsu, futa, fu, bu, pu, nu, ne, tsu, zu, ju, nyu
3	san, zan, sa, za, mitsu, mi, ta, da, so, zo, suri
4	shi, ji, su, zu, yotsu, yon, yo, fo
5	go, ko, itsutsu, itsu, i, u, ka, ga, ke, ge, faibu
6	roku, ro, mutsu, mu, me, mo, ri, ru, ro, ryu, shikkusu
7	nanatsu, nana, na, shichi, chi, te, de, yu, sebun, se, ze
8	hachi, ha, ba, pa, fa, he, be, pe, fe, yatsu, ya, eito, ei
9	kyuu, gyuu, ku, gu, kokonotsu, kokono, koko, ko, go, ki, gi, kin, kun, gin, gun, nain
10	too, to, doo, do, juu, ji, ten, den, te

Table 1. Possible readings of numbers in Japanese.

As displayed in Table 1, commonly known readings of numbers are often vocalized, devocalized, extended or shortened according to needs. For example, the number 893 can mean *yakuza*, where *za* is a vocalized version of number 3 reading *sa* (see Table 1). Some minor additions are also sometimes made, such as adding gemination (small *tsu*, making the next consonant sound doubled) or single *n*.

[2] The table was composed by analyzing real life examples of human-created *goroawase*. We also used some information displayed at http://www2u.biglobe.ne.jp/~b-jack/kouza/s-3.html, where also explanation can be found regarding some uncommon readings.

Japanese speakers may notice that Table 1 contains also sounds which are not commonly used in reference to numbers (all of them, however, have been placed in the table for particular reasons – see Footnote 2). This, if such uncommon readings were used in our system (see below), may lead to generation of *goroawase* too complicated and not very associable with what they are referring to. However, we believe that this might be actually a desirable issue in our system, as common, simple *goroawase* can be easily created by humans, while those of higher complexity are not that likely to be thought. Thus, we believe that equipping our system with such a bit of creativity should improve its usability in the eyes of potential users.

Goroawase are commonly used at school during history lessons, to help memorize important dates. For example, the year 1492 (when Columbus discovered America) can be remembered as *iyo-kuni (ga mieta)* ("wow, I can see land!"), in which the first part is a *goroawase* for the date (1-*i*, 4-*yo*, 9-*ku* and 2-*ni*). They are also quite popular in daily language, especially its genre used to communicate within modern technologies, such as cellular phones or internet. For example, the number "4649" is often used in place of *yoroshiku* (sort of a greeting; can be translated as "nice to meet you" or "please take care of it"), and "18782" can be a code of "*iya na yatsu*" ("unpleasant guy").

Goroawase have also wide commercial applicability, as they can be used to make phone numbers easier to memorize. Quite well known examples are the combinations "117 117" ("*iina iina"* – "it's good, it's good!") or "4989" ("*yoku yaku"* – "we fry well") for fried meat restaurants.

In the system proposal presented in this work we would like to address the commercial need for automatic *goroawase* generation. Below we describe our research on Japanese word plays conducted so far (Section **3**) and introduce the new idea of a *goroawase* generating system (Section **4**).

3. PUNDA Project

Started in 2007, the PUNDA Project is aimed at constructing a humor-equipped conversational

system for Japanese. During its development, we have constructed two versions of such a system: one with basic joking abilities, using humor at every third turn of conversation [1], and other with an emotiveness analysis-based timing module [2].

During our research we also developed a semantic algorithm, which allows our system to check the pun candidates' relation to the context [3]. The algorithm uses the Internet as a source of knowledge. In search engines, such as Yahoo[3], it checks the co-occurrence of pun base phrases and candidates in order to measure how closely they are related to each other. If, for instance, a base word is *katana* (a Japanese saber), and one of pun candidates generated by our system is *katta na* (I bought it), the system would check the co-occurrence of these two in the Internet, along with other candidates, and then form a ranking of all candidates according to their co-occurrence with the base word.

The algorithm mentioned above can also be used in the *goroawase* generating system.

4. *Goroawase* – computation possibilities

The features of *goroawase* (see Section **2.3**) make it an interesting subject for such fields of science as AI or NLP. Especially tools and algorithms developed in the latter seem quite usable to create a *goroawase* generating system. Thus, the more surprising is the fact that, to our knowledge, very few serious attempts of building such a system have been made so far. One existing and working engine is available on line for wide use – named "*Goroawase* generator", toward inputted set of numbers, it generates a phrase making it easier to memorize, using hand-made data bases with readings of numbers 1-22 digits long[4].

This application seems quite useful. However, in our research we are aiming to build an algorithm which would do the opposite – generate lists of easily memorable numbers towards sets of inputted keywords. Such a setup might be useful in the world of commerce, in situations where a company looks for a number that would be easy to remember and associable with its name or area of activity. The algorithm we propose below is constructed to address the need for such systems.

5. PUNDA Numbears – algorithm proposal

As a part of our main research project PUNDA (see Section **3**), we started a side project, named PUNDA Numbears. Its goal is to create a *goroawase* generating system, able to generate sets of numbers associable with inputted keywords. So far, we have designed an algorithm to do that, and currently we are working on its implementation. The algorithm outline is presented on Figure 1.

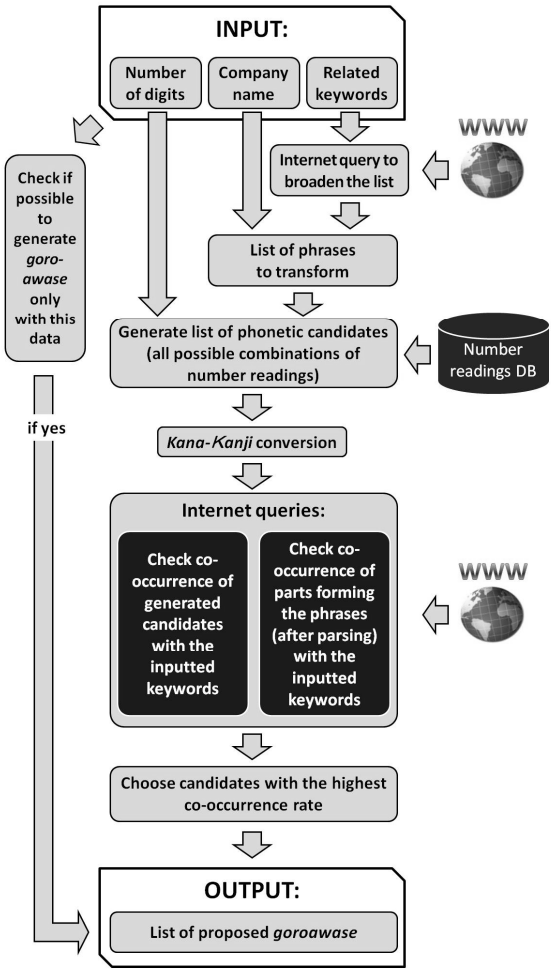

Fig.1 *Goroawase* generating system – algorithm outline.

The input for the system consists of three parts: 1) the number of digits forming a number to be generated, 2) the name of the company (or any other subject the output should be associated with) and 3) list of related keywords (such as types of

[3] www.yahoo.co.jp
[4] Available at http://seoi.net/goro/

business, names of wares or services). Words from that list are next queried in the Internet to find other words that are commonly associated with them (mostly nouns, verbs and adjectives). To do that, we are planning to use an algorithm developed by Higuchi et al. [4] in their research on conversational systems for Japanese.

Next, the whole list of phrases (with added association words) is transformed into a list of phonetic candidates. To do that, all possible combinations of number readings (according to inputted number of digits) are generated, using a data base containing information displayed in Table 1. Then, all candidates are converted into *Kanji* characters (with all options if more than one is possible).

Such a list is next processed in two ways. First, each candidate's co-occurrence with each inputted keyword is checked, and a ranking is formed to find the most related pairs. Second, each candidate is parsed using the POS and morphological analyzer MeCab [5], and its constituent parts' co-occurrence with each inputted keywords is checked, forming another ranking.

Candidates from the tops of both rankings are transformed into numbers and outputted as proposed *goroawase*. Also, if it is possible to generate *goroawase* only from the input, such propositions are also added to the list.

Below we present an example of how the system should work:

input:

-number of digits: 7

-company name: *Kurohashi*

-related keywords: *yasai* (vegetables), *kudamono* (fruits), *tabemono* (food), *yaoya* (grocery)

presumed output:

-using generated list of associations:

- 9684 014
  (*kurohashi oishii* – tasty Kurohashi: 9-*ku,* 6-*ro,* 8-*ha,*4-*shi*, 0-*o,* 1-*i*, 4-*shi*)
- 9684 910
  (*kurohashi guddo* – Good Kurohashi: 9-*ku,* 6-*ro,* 8-*ha,*4-*shi*, 9-*gu,* 10-*do*)
- 0141 210
  (*oishii fuudo* – tasty food: 0-*o,* 1-*i*, 4-*shi*, 1-*i*, 2-*fu,* 10-*do*)
- 910 2010
  (*guddo fuudo*– Good food: 9-*gu,* 10-*do*, 2-*fu,*0-*u*, 10-*do*)
  (...)

-from the inputted information:

- 9684 831
  (*kurohashi yasai*: 9-*ku,* 6-*ro,* 8-*ha,*4-*shi*, 8-*ya,* 3-*sa,* 1-*i*)
- 9684 808
  (*kurohashi yaoya*: 9-*ku,* 6-*ro,* 8-*ha,*4-*shi*, 8-*ya,* 0-*o,* 8-*ya*)
  (...)

6. Conclusion and future work

In this paper we proposed an idea of *goroawase* generating system. We briefly described the genre of *goroawase*, its major linguistic mechanisms along with their computation possibilities. We also proposed an outline of an algorithm which, if implemented, should be able to generate *goroawase* proposals towards inputted phrases.

As the work is currently in the development stage, the next step is the implementation of the idea, construction of an actual system and its evaluation, which we are hoping to do in the nearest future.

Reference:

[1] Dybala, P., Ptaszynski, M., Higuchi, S., Rzepka, R. and Araki, K.: "Humor Prevails! - Implementing a Joke Generator into a Conversational System", In: AI-08, Wobcke, W. and Zhang, M. (eds), Auckland, New Zealand. Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI) Vol. 5360, pp. 214-225, Berlin-Heidelberg, 2008.

[2] Dybala, P., Ptaszynski, M., Rzepka, R. and Araki, K.: "Humoroids – Talking Agents That Induce Positive Emotions with Humor", In: AAMAS 2009, pp. 1171-1172, Budapest, Hungary, 2009.

[3] Dybala, P., Ptaszynski, M., Rzepka, R. and Araki, K.:, "Crossing Word Borders: Towards Phrasal Pun Generation Engine", In: PACLING 2009, pp.242-247, Sapporo, Japan, 2009.

[4] Higuchi, S., Rzepka, R. and Araki, K.: "A Casual Conversation System Using Modality and Word Associations Retrieved from the Web", In: EMNLP'08, pp. 382-390, Honolulu, USA, 2008.

[5] MeCab: Yet another part-of-speech and morphological analyzer, T.Kudo, http://mecab.sourceforge.net/