Basic Idea of General Belief Retrieving Agent “GENTA”

Rafal Rzepeka, nonmember, Kenji Araki, Koji Tochinai* Regular Members

SUMMARY One of the main problems on the way to an anthropomorphic system is that machine’s knowledge differs from human’s in various ways. In our approach we endeavor to use WWW-based corpus to enable our system to discover what people generally know, think or feel about a spoken matter. Our system also inductively learns from the “user’s input - own output” pairs of sentences. In our opinion research in talking interfaces field is being lately neglected but we predict that combining natural language processing with data mining techniques will lead to creating a freely speaking system that will overgrow existing programs.

Key words: natural language interfaces, human-computer interaction, language acquisition, knowledge retrieval, inductive learning.

1. Introduction

Usually talking agents creators (from classic chatterbot ELIZA[1] to more sophisticated systems as ALICE[2]) concentrate on keywords manipulation based discourse monitoring, which is largely supported by internal databases. The outside knowledge retrieval stays in background because of a prosaic problem of the time that nowadays computers need for text data mining. We decided to start developing system, which places emphasis on the semantic quality of the system’s utterance even if extended time of output creation costs us problems with performance.

Most researches on language acquisition base on trials of copying a human child’s method of achieving an ability to communicate. In our approach we presume that a computer is different from a child, what we would like to take advantage of. Immense memory space, almost unlimited data access, etc. should be united with child’s naivety while learning. For teaching the system human’s basic reasoning and language we have to equip it with basic feelings and the will to communicate and learn. Unless machines acquire basic knowledge as “I like it” or “I dislike it”, what means to have an own opinion, talking to computers will always be machine-like. For that reason we propose a method of using the WWW sources for retrieving common beliefs and opinions about matter talked by a user and the agent. We propose a system, which abilities to talk on any topic will grow together with the Web resources. For the reason that Japanese speaking subject-unlimited system does not exist, we decided to evaluate our system by inputting the same question to GENTA and to the Tourist Guidance System for Mt. Fuji[3].

It appeared that our system was able to find 4 answers and chose one that was grammatically and logically proper without preparing any specialized data.

2. Data retrieval methods

2.1 Basic Idea

GENTA (GENeral belief reTriving Agent) uses its own searching engine to explore corpora for data. It looks not for keywords alone but for keyword frames (templates), which are chosen depending on the parsing process results.

![Fig. 1. Basic structure of GENTA system.](image)

Every user’s input is parsed by Japanese morphological analyzer ChaSen[4] then usually becomes a sentence form rule which is saved in dictionary if such input was not made before or counted as more frequent if it was. Therefore the system checks if ChaSen did not fail on classifying any words and tries to classify if it failed. The next step is creating what we call Imagination Flashes around given nouns and verbs. We plan to teach our system how to create scripts, which could be retrieved from and verified by the corpora through above-mentioned search frames for particular parts of speech in prepared combinations letting our system to place words in proper category. GENTA is to achieve the knowledge about given words only when the probability of mistake is very small. Basically, our agent is not limited to any particular topic and we presume that word which does not exist in WWW is obviously a misspell or a user’s individual invention which does not have to be necessarily understood by the system. Then the Imagination Flashes are analyzed by the Opinion Section, which tries to retrieve the Net users’ opinion about keywords and evaluate a rate of interest, which is crucial for the Utterance Section.

---

[1] The authors are with Signal Processing and Language Information Science Laboratory, Graduate School of Engineering, Hokkaido University
2.2 Processing Example

When "中国語が話せますか。" (Do you speak Chinese?) is inputted, ChaSen analyses sentence and the system simplifies the results, which decide which corpus search method will be used, what basically means counting how many times somebody wrote watashi : ha : NN[coun] : N[suff] : ga : V[t-pol] (私は中国語が話せる: I can speak Chinese) or its negation (V[t-pol:neg]). The bigger number decides what system answers if it is yes/no question. Therefore GENTA starts the data retrieval concentrated on the nouns and verbs included in input by using the sets of keyword frames (e.g., "N:ha:x:desu") that allow to guess the kind of noun and search for other information about it, e.g., "中国語は、同じ漢字圏に生活している日本人にとって親しみやすい言語です。

(Chinese language is friendlier for Japanese who also live in a country where Kanji characters are used.) All searched sentences are also analyzed morphologically and sentences structures are saved as proper Japanese which later will allow learning colloquial grammar.

(What can I do at Motosu Lake?). By using only morphological analyzed keywords our agent retrieved 2 proper and 2 erroneous answers:
①本栖湖でレジャー目的のジェットができる。(You can jet-ski for pleasure at Motosu Lake.)
②本栖湖では、本当なら五千円札に使われている富士山の姿を見ることができるポイントがあるのですか... (At Motosu lake there is a point, where you can see the view of Fuji that was used for 5000 yen note...)
③厳冬期の本栖湖では、けっこう期待できるポイントです。(It is quite promising place if you go to Motosu lake in the coldest part of the winter.)
④夏は本栖湖で水遊びができる。(During summer you can splash about in the water at Motosu lake.)

After calculating common attitude sentence ④ appeared to be the most pleasant and interesting nouns configuration and was chosen as a response. We compared above response with one done by the Mt. Fuji System which is made to narrow the spoken matter "〜そんなことになさらないですか" (What would you like to do there?).

Tourist Information System's goal is to give the best proposition to the client, GENTA's - to answer any question and learn. Interestingly, we also noticed that system chooses the most obvious sentences and even if ① could be more informative for a grown-up user GENTA answers in naive, child-like way choosing ④. This may trace us to what computers lack - awareness of obviousness.

3. Experiment and consideration

We inputted the same question to GENTA's and Mt. Fuji Information System: "本栖湖で何ができますか。"