Is It Correct? - Towards Web-Based Evaluation of Automatic Natural Language Phrase Generation

Calkin S. Montero and **Kenji Araki**
Graduate School of Information Science and Technology, Hokkaido University,
Kita 14-jo Nishi 9-chome, Kita-ku, Sapporo, 060-0814 Japan
{calkin,araki}@media.eng.hokudai.ac.jp

Abstract

This paper describes a novel approach for the automatic generation and evaluation of a *trivial dialogue phrases* database. A trivial dialogue phrase is defined as an expression used by a chatbot program as the answer of a user input. A transfer-like genetic algorithm (GA) method is used to generating the trivial dialogue phrases for the creation of a natural language generation (NLG) knowledge base. The automatic evaluation of a generated phrase is performed by producing n-grams and retrieving their frequencies from the World Wide Web (WWW). Preliminary experiments show very positive results.

1 Introduction

Natural language generation has devoted itself to studying and simulating the production of written or spoken discourse. From the *canned text* approach, in which the computer prints out a text given by a programmer, to the *template filling* approach, in which predetermined templates are filled up to produce a desired output, the applications and limitations of language generation have been widely studied. Well known applications of natural language generation can be found in human-computer conversation (HCC) systems. One of the most famous HCC systems, ELIZA (Weizenbaum, 1966), uses the template filling approach to generate the system's response to a user input. For a dialogue system, the template filling approach works well in certain situations, however due to the templates limitations, nonsense is produced easily.

In recent research Inui et al. (2003) have used a corpus-based approach to language generation. Due to its flexibility and applicability to open domain, such an approach might be considered as more robust than the template filling approach when applied to dialogue systems. In their approach, Inui et al. (2003), applied keyword matching in order to extract sample dialogues from a *dialogue corpus*, i.e., utterance-response pairs. After applying certain *transfer or exchange rules*, the sentence with maximum occurrence probability is given to the user as the system's response. Other HCC systems, e.g. Wallace (2005), have applied the corpus based approach to natural language generation in order to retrieve system's trivial dialogue responses. However, the creation of the hand crafted knowledge base, that is to say, a dialogue corpus, is a highly time consuming and hard to accomplish task[1]. Therefore we aim to automatically generate and evaluate a database of trivial dialogue phrases that could be implemented as knowledge base language generator for open domain dialogue systems, or chatbots.

In this paper, we propose the automatic generation of trivial dialogue phrases through the application of a transfer-like genetic algorithm (GA) approach. We propose as well, the automatic evaluation of the *correctness*[2] of the generated phrase using the WWW as a knowledge database. The generated database could serve as knowledge base to automatically improve publicly available chatbot[3] databases, e.g. Wallace (2005).

[1] The creation of the ALICE chatbot database (ALICE brain) has cost more that 30 researchers, over 10 years work to accomplish. http://www.alicebot.org/superbot.html http://alicebot.org/articles/wallace/dont.html

[2] Correctness implies here whether the expression is grammatically correct, and whether the expression *exists* in the Web.

[3] Computer program that simulates human conversation.

Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 5–8,
Sydney, July 2006.

2 Overview and Related Work

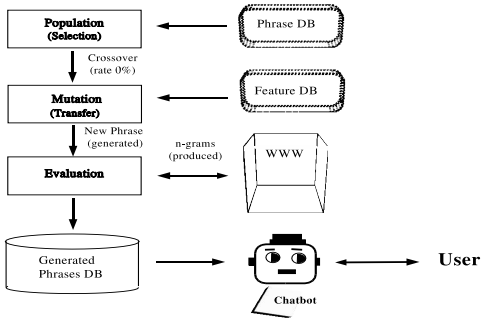

Figure 1: System Overview

We apply a GA-like transfer approach to automatically generate new trivial dialogue phrases, where each phrase is considered as a gene, and the words of the phrase represent the DNA. The transfer approach to language generation has been used by Arendse (1998), where a sentence is being *re-generated* through word substitution. Problems of erroneous grammar or ambiguity are solved by referring to a lexicon and a grammar, re-generating substitutes expressions of the original sentence, and the user deciding which one of the generated expressions is correct. Our method differs in the application of a GA-like transfer process in order to automatically insert new features on the selected original phrase and the automatic evaluation of the newly generated phrase using the WWW. We assume the automatically generated trivial phrases database is desirable as a knowledge base for open domain dialogue systems. Our system general overview is shown in Figure 1. A description of each step is given hereunder.

3 Trivial Dialogue Phrases Generation: Transfer-like GA Approach

3.1 Initial Population Selection

In the population selection process a small population of phrases are selected randomly from the Phrase DB[4]. This is a small database created beforehand. The Phrase DB was used for setting the thresholds for the evaluation of the generated phrases. It contains phrases extracted from real human-human trivial dialogues (obtained from the corpus of the University of South California (2005)) and from the hand crafted ALICE database. For the experiments this DB contained 15 trivial dialogue phrases. Some of those trivial dialogue phrases are: *do you like airplanes ?, have you have your lunch ?, I am glad you are impressed, what are your plans for the weekend ?*, and so forth. The initial population is formed by a number of phrases randomly selected between one and the total number of expressions in the database. No evaluation is performed to this initial population.

3.2 Crossover

Since the length, i.e., number of words, among the analyzed phrases differs and our algorithm does not use semantical information, in order to avoid the distortion of the original phrase, in our system the crossover rate was selected to be 0%. This is in order to ensure a language independent method. The generation of the new phrase is given solely by the mutation process explained below.

3.3 Mutation

During the mutation process, each one of the phrases of the selected initial population is mutated at a rate of $1/N$, where N is the total number of words in the phrase. The mutation is performed through a transfer process, using the Features DB. This DB contains descriptive features of different topics of human-human dialogues. The word "features" refers here to the specific part of speech used, that is, nouns, adjectives and adverbs[5]. In order to extract the descriptive features that the Feature DB contains, different human-human dialogues, (USC, 2005), were clustered by topic[6] and the most descriptive nouns, adjectives and adverbs of each topic were extracted. The word to be replaced within the original phrase is randomly selected as well as it is randomly selected the substitution feature to be used as a replacement from the Feature DB. In order to obtain a language independent system, at this stage part of speech tagging was not performed[7]. For this mutation process, the total number of possible different expressions that could be generated from a given phrase is $N^{F_{DB}}$, where the exponent F_{DB} is the total number of features in the Feature DB.

[4] In this paper DB stands for database.

[5] For the preliminary experiment this database contained 30 different features

[6] Using agglomerative clustering with the publicly available Cluto toolkit

[7] POS tagging was used when creating the Features DB. Alternatively, instead of using POS, the features might be given by hand

Total no Phrases Gen		Unnatural		Usable		Completely Natural		Precision	Recall
Accepted	Rejected	Accepted	Rejected	Accepted	Rejected	Accepted	Rejected		
80	511	36	501	18	8	26	2	0.550	0.815
Total 591		Total 537		Total 26		Total 28			

Table 3. Human Evaluation - Naturalness of the Phrases

3.4 Evaluation

In order to evaluate the correctness of the newly generated expression, we used as database the WWW. Due to its significant growth[8], the WWW has become an attractive database for different systems applications as, machine translation (Resnik and Smith, 2003), question answering (Kwok et al., 2001), commonsense retrieval (Matuszek et al., 2005), and so forth. In our approach we attempt to evaluate whether a generated phrase is correct through its frequency of appearance in the Web, i.e., the *fitness* as a function of the frequency of appearance. Since matching an entire phrase on the Web might result in very low retrieval, in some cases even non retrieval at all, we applied the sectioning of the given phrase into its respective n-grams.

3.4.1 N-Grams Production

For each one of the generated phrases to evaluate, n-grams are produced. The n-grams used are bigram, trigram, and quadrigram. Their frequency of appearance on the Web (using Google search engine) is searched and ranked. For each n-gram, thresholds have been established[9]. A phrase is evaluated according to the following algorithm[10]:

`if` $\alpha < NgramFreq < \theta$, `then` $Ngram$ "weakly accepted"
`elsif` $NgramFreq > \theta$, `then` $Ngram$ "accepted"
`else` $Ngram$ "rejected"

where, α and θ are thresholds that vary according to the n-gram type, and $NgramFreq$ is the frequency, or number of hits, returned by the search engine for a given n-gram. Table 1 shows some of the n-grams produced for the generated phrase "what are your plans for the game?" The frequency of each n-gram is also shown along with the system evaluation. The phrase was evaluated as *accepted* since none of the n-grams produced was *rejected*.

	N-Gram	Frequency (hits)	System Eval.
Bigram	what:are	213000000	accepted
Trigram	your:plans:for	116000	accepted
Quadrigram	plans:for:the:game	958	accepted

Table 1. N-Grams Produced for:
"what are your plans for the game?"

4 Preliminary Experiments and Results

The system was setup to perform 150 generations[11]. Table 2 contains the results. There were 591 different phrases generated, from which 80 were evaluated as "accepted", and the rest 511 were rejected by the system.

Total Generations	150
Total Generated Phrases	591
Accepted	80
Rejected	511

Table 2. Results for 150 Generations

As part of the preliminary experiment, the generated phrases were evaluated by a native English speaker in order to determine their "naturalness". The human evaluation of the generated phrases was performed under the criterion of the following categories:

a) Unnatural: a phrase that would not be used during a conversation.

b) Usable: a phrase that could be used during a conversation,even though it is not a common phrase.

c) Completely Natural: a phrase that might be commonly used during a conversation.

The results of the human evaluation are shown in Table 3. In this evaluation, 26 out of the 80 phrases "accepted" by the system were considered "completely natural", and 18 out of the 80 "accepted" were considered "usable", for a total of 44 *well-generated* phrases[12]. On the other hand, the system mis-evaluation is observed mostly within the "accepted" phrases, i.e., 36 out of 80 "accepted" were "unnatural", whereas within the "rejected" phrases only 8 out of 511 were considered "usable" and 2 out of 511 were considered "completely natural", which affected negatively the pre-

[8] As for 1998, according to Lawrence and Giles (1999) the "surface Web" consisted of approximately 2.5 billion documents. As for January 2005, according to Gulli and Signorini (2005),the size of indexable Web had become approximately 11.5 billion pages

[9] The tuning of the thresholds of each n-gram type was preformed using the phrases of the Phrase DB

[10] The evaluation "weakly accepted" has been designed to reflect n-grams whose appearance on the Web is significant even though they are rarely used. In the experiment they were treated as *accepted*.

[11] Processing time: 20 hours 13 minutes. The Web search results are as for March 2006

[12] Phrases that could be used during a conversation

Original Phrase	Generated Phrase
what are your plans for the weekend ?	**Completely Natural**
	what are your plans for the game ?
	Usable
	what are your friends for the weekend ?
	Unnatural
	what are your plans for the visitation ?

Table 4. Examples of Generated Phrases

cision of the system.

In order to obtain a statistical view of the system's performance, the metrics of recall, (*R*), and precision, (*P*), were calculated according to (*A* stands for "Accepted", from Table 3):

$$R = \frac{Usable(A)+CompletelyNatural(A)}{UsableTotal+CompletelyNaturalTotal}$$

$$P = \frac{Usable(A)+CompletelyNatural(A)}{Unnatural(A)+Usable(A)+CompletelyNatural(A)}$$

Table 4 shows the system output, i.e., phrases generated and evaluated as "accepted" by the system, for the original phrase "what are your plans for the weekend ?" According with the criterion shown above, the generated phrases were evaluated by a user to determine their naturalness - applicability to dialogue.

4.1 Discussion

Recall is the rate of the *well-generated* phrases given as "accepted" by the system divided by the *total number* of well-generated phrases. This is a measure of the coverage of the system in terms of the well-generated phrases. On the other hand, the precision rates the well-generated phrases divided by the total number of "accepted" phrases. The precision is a measure of the correctness of the system in terms of the evaluation of the phrases. For this experiment the recall of the system was 0.815, i.e., 81.5% of the total number of well-generated phrases where correctly selected, however this implied a trade-off with the precision, which was compromised by the system's wide coverage.

An influential factor in the system precision and recall is the selection of new features to be used during the mutation process. This is because the insertion of a new feature gives rise to a totally new phrase that might not be related to the original one. In the same tradition, a decisive factor in the evaluation of a well-generated phrase is the constantly changing information available on the Web. This fact rises thoughts of the application of *variable* threshold for evaluation. Even though the system leaves room for improvement, its successful implementation has been confirmed.

5 Conclusions and Future Directions

We presented an automatic trivial dialogue phrases generator system. The generated phrases are automatically evaluated using the frequency hits of the n-grams correspondent to the analyzed phrase. However improvements could be made in the evaluation process, preliminary experiments showed a promising successful implementation. We plan to work toward the application of the obtained database of trivial phrases to open domain dialogue systems.

References

Bernth Arendse. 1998. Easyenglish: Preprocessing for MT. In *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98)*, pages 30–41.

Antonio Gulli and Alessio Signorini. 2005. The indexable web is more than 11.5 billion pages. In *In Proceedings of 14th International World Wide Web Conference*, pages 902–903.

Nobuo Inui, Takuya Koiso, Junpei Nakamura, and Yoshiyuki Kotani. 2003. Fully corpus-based natural language dialogue system. In *Natural Language Generation in Spoken and Written Dialogue, AAAI Spring Symposium*.

Cody Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. *ACM Trans. Inf. Syst.*, 19(3):242–262.

Steve Lawrence and Lee Giles. 1999. Accessibility of information on the web. *Nature*, 400(107-109).

Cynthia Matuszek, Michael Witbrock, Robert C. Kahlert, John Cabral, Dave Schneider, Purvesh Shah, and Doug Lenat. 2005. Searching for common sense: Populating cyc(tm) from the web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380.

University of South California USC. 2005. Dialogue diversity corpus. http://www-rcf.usc.edu/˜billmann/diversity/DDivers-site.htm.

Richard Wallace. 2005. A.l.i.c.e. artificial intelligence foundation. http://www.alicebot.org.

Joseph Weizenbaum. 1966. Elizaa computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.