Genetic Algorithm-like Approach to Natural Language Phrase Generation - Case of Study Spanish

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 an approach to natural language generation for Spanish language. The proposed system automatically generates and evaluates Spanish trivial phrases using a genetic algorithm (GA)-like transfer approach and n-gram frequency information obtained from the Web. The experiment showed the encouraging results of a 79.06% user understandability of the good phrases generated by the system.
Keywords: GA-like transfer approach, trivial phrases, Web frequency information, Spanish

1 Introduction

The problematic domain of human-computer conversation (HCC) has been of particular interest to NLP-researchers, prompting them to develop conversational systems that range from applications to goal specific computer spoken dialogue, e.g., Jupiter, providing a telephone-based conversational interface for international weather [1], airline travel information systems [2], restaurant guides [3], telephone interfaces to emails or calendars [4], and so forth, to attempts to simulate human trivial dialogue - chat - e.g. ELIZA [5]. From the *canned* text approach to the *template filling* approach, different methods have been proposed to battle the lack of computer understanding when it is talking with a user.

In recent research Inui et al. [6] have used a corpus based approach to language generation for dialogue system. 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. Other HCC systems, e.g. Wallace [7], Carpenter [8] have applied as well a corpus based approach to natural language generation in order to retrieve system's trivial dialogue responses to user inputs.

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]. In this paper we describe a work in progress 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 genetic algorithm (GA)-like transfer 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). The results obtained after applying the algorithm to Spanish language are shown.

2 System Overview and Method

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 [9], 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

[1] The creation of the ALICE chatbot database (ALICE brain) has cost more that 30 researchers, over 10 years work to accomplish. The Jabberwacky database is being developed since 1988 (on the Web since 1997) http://www.alicebot.org/superbot.html http://alicebot.org/articles/wallace/dont.html http://feeling.jabberwacky.com:8081/j2about

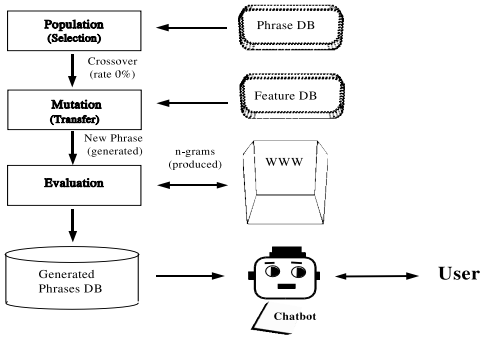

Fig. 1. System Overview

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 Fig. 1. A description of each step is given hereunder.

2.1 Initial Population Selection

In the population selection process a small population of phrases is selected randomly from the Phrase DB[2]. This is a small database created beforehand. It contains phrases used during real human-human trivial dialogues. For the experiments this DB contained 20 trivial dialogue phrases. Some of those trivial dialogue phrases are: *qué tipo de música te gusta ? (what kind of music do you like?), qué vas a comer hoy ? (what are you going to eat today?)* and 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.

2.2 Crossover

Since our algorithm does not use syntactic information (part of speech tagging), 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.

2.3 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.The word "features" refers here to the specific part of speech used, that is, nouns, adjectives and adverbs. For the experiment this database contained 100 different features. The word to be replaced within the original phrase is randomly selected as well as it is randomly selected the 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. The one-hundred features used were selected randomly from the most frequents words in Spanish from [10][3]

[2] In this paper DB stands for database.
[3] as appear in http://www.um.es/lacell/proyectos/dfe/

Table 1. Human Evaluation: Understandability of the Phrases

	Human Evaluation (Semantics)			Human Evaluation (Grammar)		
System Evaluation	5	3	1	5	3	1
Good [43]	18.60% [8/43]	60.46% [26/43]	20.93%[9/43]	30.23% [13/43]	32.56% [14/43]	37.21% [16/43]
Usable [340]	12.06% [41/340]	49.41% [168/340]	38.53% [131/340]	13.82% [47/340]	39.12% [133/340]	47.06% [160/340]
Rejected [1384]	0.66% [9/1384]	4.12% [65/1384]	94.63% [1310/1384]	0.79% [11/1384]	4.06% [56/1384]	95.15% [1317/1384]

2.4 Evaluation

In order to evaluate the newly generated expression, we used as database the WWW. Due to its significant growth the WWW has become an attractive database for different systems applications as, machine translation [11], question answering [12], commonsense retrieval [13], and so forth. In our approach we attempt to evaluate whether a generated phrase is *correct* through the frequency of appearance of its n-grams 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. The n-grams frequency of appearance on the Web (using Google search engine) is searched and ranked. A phrase is evaluated according to the following algorithm:
`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. The number of n-grams created for a given phrase is automatically determined by the system and it varies according to the length of the phrase. The system evaluates a phrase as "good" if all of its n-grams were labeled "accepted" by the system, it is to say, all of the n-grams are above the given threshold. If for a given phrase there is at most *one rejected n-gram* or *one weakly accepted n-gram*, the phrase is evaluated as "usable". The rest are "rejected".

3 Experiments and Results

The system was setup to perform 1,000 generations.There were 1,767 different phrases generated, from which 43 were evaluated as "good", 340 were evaluated as "usable" and the rest 1,384 were rejected by the system.

As part of the experiment, the generated phrases were evaluated by a native Spanish speaker in order to determine their "understandability". By understandability here we refer to the semantic information contain by the phrase, that is to say, how well it expresses information to the user. We argue that one of the characteristics of human-chat is the ability to express semantic information within a given context. This implies that a given phrase does not necessarily have to be grammatically correct in order to be understood and used. Plenty of examples are seen in the chatting rooms of the Web, in any language. The human evaluation of the generated phrases was performed under the criterion of the following categories:
a) Grammatical correctness: ranging from 1 to 5, where 1 is incorrect and 5 is completely correct.
b) Semantic correctness: evaluates the ability of the phrase to convey information. Asking the question: is the meaning of the phrase understandable? It ranges from 1 to 5, where 1 is no-understandable, and 5 is completely understandable.

The results of the human evaluation are shown in Table 1. According to the user evaluation of the semantics for the "usable phrases" evaluated by the system, 61.47% (considering the pharses evaluated from 3 to 5) of those phrases were semantically understood by the user. The understandability rose considerably, to 79.06%, for the "good phrases". The porcent of correct grammar for the "good phrases" was around 62.79% while for the "usable phrases" ranked around 52.94%. The system was able to correctly reject a semantically and grammatically wrong phrase in around 95% of the cases. However partially, those results reinforce our argument regarding the prevalecency of the partial semantics over complete correct grammar for trivial dialogue communication.

3.1 Discussion

The nature of the Spanish language, e.g. article, noun, genre and number agreement in a sentence, made it highly difficult to generate new good phrases during this experiment. Therefore, as a measure to diminish this negative effect and to give more opportunity to the generation of good phrases, the nouns used in the Feature DB were given in two versions: with its correct article, e.g. la, las, el, etc, and without article. Even though the generation rate of good phrases was low, there was a 79% of agreement with the user understanding of the phrases. Although there are not conclusive results, from this experiment a hint is revealed on the importance of the meaning of a given phrase beyond *completely* correct grammar. However there is still plenty of room for research, the results of this experiment are promising.

4 Conclusions and Future Work

In this paper the automatic generation of trivial dialogue phrases was shown through the application of a GA-like transfer approach having as case of study Spanish language. The automatic evaluation of a generated Spanish phrase using the WWW as a knowledge database was analyzed and hints on the prevalence of understandability over completely correct grammar were seen in the user evaluation. As on going work, room is left to survey other users opinions as for exploring applying the system to other languages.

Acknowledgment

This work has been partially supported by a grant of the Ministry of Internal Affairs and Communications Strategic Information and Communications R&D Promotion Program (SCOPE).

References

1. Zue Victor, Seneff Stephanie, Glass James, Polifroni Joseph, Pao Christine, Hazen Timothy J., and Hetherington Lee. Jupiter: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8:85–96, 2000.
2. Harry Bratt, John Dowding, and Kate Hunicke-Smith. The sri telephone-based atis system. In *Proceedings of the ARPA Spoken Language System Technology Workshop*, pages 22–25, 1995.
3. Raymond Lau, Giovanni Flammia, Christine Pao, and Victor Zue. Webgalaxy: Beyond point and click - a conversational interface to a browser. In *Proceedings of the 6th International WWW Conference*, pages 119–128, 1997.
4. The University of Texas at Austin. Smartvoice. http://www.utexas.edu/its/smartvoice/.
5. Joseph Weizenbaum. Eliza a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, 1966.
6. Nobuo Inui, Takuya Koiso, Junpei Nakamura, and Yoshiyuki Kotani. Fully corpus-based natural language dialogue system. In *Natural Language Generation in Spoken and Written Dialogue, AAAI Spring Symposium*, 2003.
7. Richard Wallace. A.l.i.c.e. artificial intelligence foundation, 2005. http://www.alicebot.org.
8. Rollo Carpenter. Jabberwacky. learning artificial intelligence. http://www.jabberwacky.com http://www.icogno.com/.
9. Bernth Arendse. Easyenglish: Preprocessing for mt. In *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98)*, pages 30–41, 1998.
10. Ramón Almela, Pascual Cantos, Aquilino Sánchez, Ramón Sarmiento, and Moisés Almela. *Frecuencias del Espaȕol Diccionario y estudios léxicos y morfológicos*. Editorial Universitas, S.A, –.
11. Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, 2003.
12. Cody Kwok, Oren Etzioni, and Daniel S. Weld. Scaling question answering to the web. *ACM Trans. Inf. Syst.*, 19(3):242–262, 2001.
13. Cynthia Matuszek, Michael Witbrock, Robert C. Kahlert, John Cabral, Dave Schneider, Purvesh Shah, and Doug Lenat. Searching for common sense: Populating cyc(tm) from the web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, 2005.