NonI: A Non-Intelligent Conversation Simulator

Jason L. Hutchens

September 4, 1998

THE LOEBNER CONTEST

An instantiation of the Turing Test, which I've entered the last three years.

In 1999, a new prize of $25,000 in US dollars is up for grabs, for any program which fools the judges more than 50% of the time into believing it is human.

The judges only get twelve minutes to evaluate each confederate. Typing and thinking delays mean that, on average, they only get to say twenty things.

This justifies taking things a little bit more seriously!

PREVIOUS ENTRIES

My first entry was named *HeX*. It was a simple Eliza-clone, and managed to win the contest in 1996.

My second entry was co-authoured with Bruce Cooper. It was named *SEPO*, and came second in the 1997 contest. It is an interpreted programming language, tailored for string processing and pattern matching.

My third entry was *MegaHAL*, which uses a simple Markov chain to generate sentences which are linked to the user's input by the words with the highest information content. It came second in 1998.

MEGAHAL

I was surprised that MegaHAL was deemed to be more human than four other programs.

A problem with sentences generated by Markov chains is that, the further along the sentence you go, the less relevance the current word has to the beginning words.

More constraints are needed.

Generation is done by fleshing a few keywords out into a sentence. The nature of Markov models makes it hard to generate sentences which must contain two or more specified words.

A NEW APPROACH

Take the user's input, and UpWrite it to the "concept level". This converts the sentence to some abstract root form.

During the UpWrite, extract variables from the sentence which may differ between users, such as the name of the user.

A contextual model takes the sequence of "concepts" thus far, and determines a probability distribution over possible reply "concepts".

A reply "concept" is selected, and DownWritten to a sentence, which is displayed to the user.

THE UPWRITE

We need to extract "meaning" from the sentence. In a very basic way, this can be done by finding word pairs with low mutual information.

We need a model which can give us the probability of a word appearing in a sentence, and the probability of a pair of words appearing, regardless of their distance of separation.

The word pair with the lowest mutual information will be taken as the "concept" of the sentence.

FINDING CONCEPTS

The word-pair in the sentence with the highest mutual-information is used. If no such pair exists, the special SS-SE pair is used.

Anatomy DESCRIBES the structure and organisation of living CREATURES.

The CARROT is a VEGETABLE.

The earth is the THIRD PLANET from the sun.

An ALPHABET is an ordered series of LETTERS used in language.

This suggests a structure for the model...

A RECURSIVE MODEL

The first model I tried stored a probability distribution over all words which could occur between a pair, where a pair may be separated by any number of words.

This model didn't have enough constraints. Infinitely long sentences were possible.

The sentences satisfied long-distance dependencies:

```
A sword is an offensive weapon designed
chiefly for the sale and consumption of
deep unconsciousness.
```

```
Dance is a person who suffers from the
carcass of movement.
```

THE NEW MODEL

To introduce more constraints, I changed the model so that the word occurring within the pair must be adjacent to at least one of the elements in the pair.

This is like having many first-order Markov chains, each one constrained by a word which occurs somewhere earlier (or later) in the sentence.

The model also incorporates all of the probabilities needed to calculate the mutual information.

SOME EXAMPLES

A symbol is a suspended brass disk which represents something else.

A monkey is a tree native to Australia where it is called the gum tree dwelling primate.

The kiwi is a local mode of pronunciation in New Zealand.

The kangaroo is a marsupial mammal found in a marsupial mammal.

The nose is an English naturalist. He published his theory of smell.

Leonardo da Vinci was an Italian artist and expert in Kung Fu who popularised the martial arts in unpublished note books.

LATEST MODEL

Each model in a family of second-order Markov models is constrained by a different "future symbol".

This means that, when the model is used generatively, we can generate towards a target.

We can generate sentences which are required to contain any number of words by specifying them as "waypoints".

The program is not yet complete, so I have no examples to give.

THE DOWNWRITE

Word-pairs are extracted from the user input. A second-order Markov chain is used to predict what the reply word-pair should be, given the previous reply word-pair and the current input word-pair.

The reply word-pair is downwritten using the model in a generative way.

Many downwrites are performed, and a second-order Markov chain is used to select the "best" one, by choosing a downwrite with a typical level of information.

CONCLUSION

The model incorporates long-distance dependencies. Generation of sentences containing any number of keywords is easy. Calculation of the mutual information of word-pairs is simple.

The contest is in January 1999. Between now and then, NonI will be made available on the Web. Web lusers will train the program by talking to it.

A lot of improvements could be made (upwriting to word-class, for instance). Suggestions are more than welcome. Offers to help with coding are even more welcome!