SEMANTIC COOCCURRENCE NETWORKS

Dan W. Higinbotham
The University of Texas at Austin
and
Executive Communication Systems
455 N. University, Suite 202
Provo, Utah 84601

One of the thorniest problems in Machine Translation is lexical ambiguity. There are many examples like the following:

The star was utmost on the astronomer's mind.
The star was utmost on the director's mind.

Although either instance of 'star' could have the other meaning in some context, the most natural reading would make 'star' in the first sentence a celestial object, and 'star' in the second sentence a human performer. Preference semantics approaches usually don't handle this situation, since the word that triggers the appropriate meaning is not part of a predication with argument preferences. In fact, the trigger word can occur almost arbitrarily far from the ambiguous word. Many systems simply default to the most frequently occurring meaning in such cases as a best guess.

Recent work by Ken Church provides hope that perhaps statistical cooccurrence methods may be of some value in disambiguating such examples. The idea of this paper is to use text to gather cooccurrence statistics on source words that commonly appear near a word when it is used in its various senses. This data is then fed to a neural network, which generalizes the information so that a most likely sense is hypothesized even for contexts that were not among the original data.

This paper will describe three groups of experiments. The first discusses *a* neural network that correctly categorized senses of a small set of ambiguous words. The second is an application of the neural network idea to senses of the word 'bank' as used in the text of definitions in the Longman Dictionary of Contemporary English. The third is an application of the idea to a quarter million words of parallel French and English text.

NEURAL COOCCURRENCE

The idea of this experiment was to see if a neural network could correctly categorize senses of ambiguous words. The input words and general categories of intended senses were as follows:

bank	money	water
bail	container	money
bat	animal	sports
bridge	cards	water

calf	animal	bodypart
chest	bodypart	container
fence	container	sports
file	office	tool
palm	bodypart	plant
pen	container	office
pitcher	container	sports
poker	cards	tool
pool	sports	water
spade	cards	tool
squash	plant	sports

The network was a feed-forward neural network. Such a network is a connected directed acyclic graph; it is composed of nodes and one-way connections between pairs of those nodes. Each node in the network has an associated real number called its activation level, and another real number called its bias. Each connection is also associated with a real number called its weight. The activation of each node changes once each cycle, and is calculated based on the following equation:

$$a_i = \Phi \left(b_i + \Sigma a_j * w_{ji} \right)$$

where a_i is the activation of node i
a_j the activation of node j
b_i the bias of node i
w_{ji} the weight of the connection from node j to node i

and

$$\Phi (x) = \frac{1}{1 + e^{-x}}$$

This particular network was composed of three sets of nodes, namely input nodes, hidden nodes, and output nodes. Each input node had a directed connection to each hidden node, and each hidden node had a directed connection to each output node. The network had 16 words as input units, 8 hidden units, and 32 output units. Each input unit corresponded to one of the 16 words. There was one output unit for each sense of each word. The network was set up using the bp (back propagation) program in the Explorations in Parallel Distributed Processing (PDP) software of McClelland and Rumelhart. The biases and weights in the network were trained according to the back propagation formulas. In these formulas, each non-input unit has an error term calculated. The error terms for output units are simply

$$E_i = t_i - a_i$$

namely, the target activation minus the actual activation.

The error term for each hidden unit depends on activations and error terms of all *of* the output units the hidden unit is connected to, as follows:

$$E_i = \sum_j W_{ij} * D_i$$

$$\text{where } D_j = E_j * a_j * (1-a_j)$$

Weights and biases are changed based on these values and constants which control the gradual descension of the network into a stable state which has learned the patterns presented. The change in each bias and weight is stored for use the next time they are calculated; these are

$$\delta w_{ij} = \varepsilon_{ij} * D_i * a_j + \mu * \delta w_{ij}$$

$$\delta b_i = \beta_i * D_i + \mu * \delta b_i$$

The constants are p (for momentum) and ε_{ij} for weight learning rate and β_i for bias learning rate. Each weight is then modified by adding the δw_{ij} term, and each bias is modified by adding the δb_i term. Each weight and bias in the network can be modified once each time *a* training pattern is presented, or once after the whole set of training patterns has been presented. For this test, weights were modified after each training pattern was presented, and the parameters were μ=.9, ε= . 5 for connections from hidden units to output units, ε=.08 for connections from input to hidden units, β=.5 for output units, and β=.08 for hidden units.

The following pairs sharing a sense in the same category were presented as input to the network:

pool bank	(water)	pool pitcher	(sports)
pool bridge	(water)	pool fence	(sports)
bank bridge	(water)	pool squash	(sports)
bank bail	(money)	pool bat	(sports)
bank stock	(money)	pitcher squash	(sports)
bail stock	(money)	pitcher bat	(sports)
stock calf	(animal)	fence squash	(sports)
stock bat	(animal)	fence bat	(sports)
calf bat	(animal)	squash bat	(sports)
chest calf	(bodypart)	bail chest	(container)
chest palm	(bodypart)	bail pitcher	(container)
palm calf	(bodypart)	bail fence	(container)
squash palm	(plant)	bail pen	(container)
file pen	(office)	chest pitcher	(container)
file spade	(tool)	chest fence	(container)
file poker	(tool)	chest pen	(container)
bridge spade	(cards)	pitcher fence	(container)
bridge poker	(cards)	pitcher pen	(container)

For each training pattern, the activations of the input units corresponding to the two input words were set to 1, and the target activations of the correct output senses were set to 1, but the target activations of the incorrect senses were set to 0. The error terms of all output units which were not senses of the input words were set to 0. The total sum of squares is a measure of how well the network has learned the input patterns; it is simply the sum of the squares of the error terms of the output units. The network in this test was trained until the total sum of squares fell below 0.4. All of these parameters, as well as the architecture of the network, were specified in PDF network and pattern files.

After the network was trained, the values of the hidden units were examined for each of the training patterns. The activations for the hidden units of each pattern (rounded to 0 if less than .5, and to 1 otherwise) were as follows:

```
11010000  pool bank   (water)
11010000  pool bridge (water)
11010000  bank bridge (water)

10000010  bank bail  (money)
10110110  bank stock (money)
10100010  bail stock (money)

00110000  stock calf (animal)
00110110  stock bat  (animal)
00111100  calf  bat  (animal)

10101000  chest calf (bodypart)
10101000  chest palm (bodypart)
00101000  palm  calf (bodypart)

01000000  squash palm (plant)

01011101 pool pitcher    (sports)
01011101 pool fence      (sports)
01010100 pool squash     (sports)
01010100 pool bat        (sports)
01010110 pitcher squash  (sports)
00010101 pitcher bat     (sports)
01010110 fence squash    (sports)
00010101 fence bat       (sports)
01010100 squash bat      (sports)

10101011  bail chest      (container)
10101111  bail pitcher    (container)
10001111  bail fence      (container)
10101111  bail pen        (container)
10101111  chest pitcher   (container)
10001111  chest fence     (container)
10101111  chest pen       (container)
10101111  pitcher fence   (container)
10001111  pitcher pen     (container)
```

```
00000111 file pen      (office)

00010011 file spade    (tool)
00010011 file poker    (tool)

10110000 bridge spade  (cards)
10110000 bridge poker  (cards)
```

From this data, it appears that the network generalized to representing not specifically the pair of inputs, but rather the common class that both of their senses belong to when they appear together. The classes could be summarized as follows:

office	00000111
tool	00010011
animal	0011???0
sports	0?01?1??
plant	01000000
bodypart	?0101000
container	10?01?11
money	10??0010
cards	10110000
water	11010000

In other words, the network set up what could be considered a binary feature representation of the underlying cooccurrence classes.

TO THE BANK

The Longman Dictionary of Contemporary English was searched for occurrences of the word 'bank' within the texts of definitions (parenthetical material was not considered).

The following are words that cooccur three or more times with 'bank' in its financial meaning within those definitions:

account book business can central certain cheque close door give interest money order paper particular pay people person print public put record room state sum supply system take various

The following words cooccur at least three times with the geographical meaning of 'bank':

built earth ground high lake overflow river rock sand sea stone stream underwater water wide

A program and parameters similar to the one above was used, with eight hidden units; one input unit for each context word and one for 'bank'; and one output unit for each word except 'bank', which had one for financial 'bank' and one for geographical 'bank'. Each training pattern consisted of a pair of input words, namely 'bank' and one of the above context words; in the target patterns, the context word and the correct sense of 'bank' had target activations of 1, the incorrect sense of 'bank' had target activation of 0, and the error terms of other output units were set to 0.

Each definition containing the word 'bank' was then presented to the network as input. For each definition, an input unit was given an activation of *I* if the unit corresponded to a word which occurred in the definition (possibly more than once); otherwise it was given an activation of 0. The activation levels of the output units for the senses of 'bank' were then compared, and the one with the highest activation was chosen to be the proper meaning in that context.

Of 97 definitions including the word 'bank', 32 referred to a geographical bank. Of these 97 definitions, the network got the wrong meaning of 'bank' in only one case, namely the definition

> "a deep bank or mass of snow formed by the wind"

which includes none of the context words above.

The experiment was tried again, using only the first 76 definitions as input for training pairs. Only words which cooccurred three or more times with 'bank' in these 76 definitions were used as companions for 'bank' in training pairs. The words 'door', 'particular', and 'room' were no longer paired with financial bank, and the words 'ground', 'high', and 'underwater' were no longer paired with geographical bank. The remaining 21 definitions were then presented as input, and the network failed in 2 cases, one being the one above, and the other being

> "a bridge consisting of a high tower on each bank connected by lengths of steel rail from which a flat carriage level with the ground hangs"

The words 'high' and 'ground' triggered correct resolution in the first case, but were not known as trigger words in the second. The network therefore succeeded in over 90% of the cases of previously unseen definitions.

RAW TEXT

Another experiment was conducted that was based on approximately a quarter million words of English and corresponding French text, which contained a variety of government and non-government documents. The text was divided into over 6000 sections, each of which contained from one to seven sentences. Function words were removed, and remaining words were reduced to base form. A bilingual dictionary was created that showed the most common French translations for the 9103 English base words that occurred in the corpus. The text was searched for pairs of English words that cooccurred three or more times; if the pair of English words mapped to the same pair of French translations in 85% of their cooccurrences, the cooccurrence was deemed to be significant. The list of pairs and their translations was used to create a new bilingual dictionary of 2855 English words mapping to 2462 French words.

The PDP software was rewritten to handle a larger network, but used the same formulas and parameters. The network used had one input unit for each English word, 54 hidden units, and one output unit for each French word. The epsilon parameters and momentum parameter were the same as above. Weights were modified after each training pattern, and the order of patterns to be presented was permuted each time before the whole set of patterns was presented.

Each cooccurring pair of English words was presented as a training pattern; the French words they mapped to were given a target activation of 1, and other possible French translations for the two English words were given target activations of 0. The error terms of other French words were set to 0. The 20,738 patterns were each presented to the network 174 times.

The network was tested on the five English words 'articles', 'committee', 'company', 'major', and 'office.' 'Articles' was considered a base form, since it occurs as a key word in the Longman Dictionary of Contemporary English. Each word will be listed with its translations and common cooccurrences, the total number of instances of the word in the corpus, the percentage of cases that could be achieved simply by choosing the most frequent translation, and the percentage achieved by the network.

articles	articles	aid, appropriate, attendance, base, committee, confidential, council, debate,only, other, parliament, parliamentary, sign, substance, treaty, unofficial
	statut	accordance, activity, board, company, directive, dividend, entrust, form, limit, ordinary, possibility, proposal, protection, provisions, register, related, second, shares, status, statutory

32 instances, 56% translated 'statut', network achieved 87%

committee	commission	apply, articles, brief, cassette, confidential, confidentiality, delegation, enable, experience, head, last, open, option, parliament, political, sign, submission,substance, superior, unofficial
	committee	trust
	comité	alternance, centre, complementary, concrete, consultative, decision-making, desirable, division, employment, especially, foreigner, generally, impact, importance, instrument, integration, language, least, needs, position, principle, problem, qualifications, recommendation, session, situation, social, standing, technology, town, unemployment

159 instances, 54% translated 'comité', network achieved 60%

company	compagnie	---
	entreprise	acceptable, common, compensate, complexity, conditions, crisis, difficulty, discourage, education, expose, handicap, key, knowledge, lighting, manpower, objective, optical, organize, permit, population, possibly, potential, principle, productive, requirement, responsibility, signal, skill, sound, strengthen, technological, transmission, treat, type, venture
	société	articles, corresponding, debtor, directive, entrust, explicitly, governing, indirectly, issue, legally, list, meeting, network, normal, bureau, orders, payable, register, regulation, statutory, structural, third, three

166 instances, 62% translated 'entreprise', network achieved 66%

major	grand	budgetary, machine, quality
	gros	appliance, asset, domestic
	important	lighting, steel
	majeur	alternate, court
	principal	--

114 instances, 43% translated 'grand', network achieved 47%

office	bureau	abroad, communication, hour, manual, migrant, transport
	office	enterprise, federation, finance

37 instances, 54% translated 'bureau', network achieved 62%

These examples are not nearly as clear cut as the 'bank' example above, partly because there are gradations of meaning and areas of overlap in the French translations, and some of the cooccurrence pairs discovered are more accidental than meaningful.

A further experiment was conducted that used training pairs based only on the first 80% of the parallel texts. The new network had 2648 input units, 52 hidden units, and 2299 output units. The following results were obtained for the previous five words running this new network on the remaining 20% of the parallel texts:

articles
  5 instances, 0% translated statut, network achieved 100%

committee
  10 instances, 10% translated comité, network achieved 50%

company
  68 instances, 76% translated entreprise, network achieved 44%

major
  40 instances, 37% translated major, network achieved 37%

office
  0 instances

Of the words corresponding to input units of the network, slightly over 900 had more than one meaning in the original dictionary. The network achieved correct resolution of 64.6% of the instances of ambiguity of the 900-plus words. Using the most frequent sense of each of the words, based on the first 80% of the parallel texts, resolves the instances of ambiguity of these words correctly in 69.2% of the cases. By counting the number of trigger words in the same text section as an instance of an ambiguous word, and choosing the sense with the most trigger words (picking the sense with the highest frequency in case of a tie), 76.9% of the ambiguous instances of these words were resolved correctly.

General vocabulary words often present the greatest ambiguity problems, because none of the word senses may be particular to a given sublanguage domain. Domain-specific knowledge bases are of little help in resolving this kind of ambiguity, and often revert to the best guess strategy in these cases.

The experiments discussed here give hope that cooccurrence statistics and sufficiently trained neural networks may be able to improve the resolution of ambiguity in these most difficult cases.

CONCLUSION

Approaches to lexical ambiguity that depend on selectional restrictions (such as Preference Semantics and its derivatives) often leave *a* class of problems that are very difficult to resolve. It appears that neural networks based on statistical cooccurrence may provide some hope for resolving lexical ambiguity in such cases.

The first experiment reported in this paper showed that neural networks are capable of generalizing from cooccurrence examples, and can apparently develop feature representations for hidden classes of cooccurrence. The second example shows a case where the presentation of cooccurrence pairs found in real text can reproduce the correct senses of the word in new text with a rate of over 90% accuracy. The third experiment shows that in cases where there appears to be sufficent cooccurrence training data, a neural network is capable of providing better ambiguity resolution than the best guess strategy.

DIRECTIONS FOR FURTHER RESEARCH

There are many parameters involved in the neural network and in preparation of the data which could be altered. For example, the size of context window, the minimum number of cooccurrences and the percentage of required matches to posit training pairs, could be varied. The number of hidden units, momentum and learning rate parameters, the length of training, or even the network architecture could also be modified. Different input strengths could be assigned to context words depending on their distance from the ambiguous word. Marking the parallel texts with word class tags, and limiting words so that they would be triggers only when they were used in a certain word class, could also improve network performance.

ACKNOWLEDGEMENTS

I would like to thank the Humanities Research Center at Brigham Young University for providing access to a machine readable copy of the Longman Dictionary of Contemporary English, and to Alan Melby for providing the parallel English and French texts.

REFERENCES

Church, Kenneth, and William Gale, Patrick Hanks, and Donald Hindle. "Parsing, Word Associations, and Typical Predicate-Argument Relations", Parsing Technologies Conference Proceedings, Carnegie Mellon University, 1989.

McClelland, James L. and David E. Rumelhart, Explorations in Parallel Distributed Processing, MIT Press, 1988.