OLE TOGEBY

Translation of Prepositions by Neural Networks

Abstract

Translation of prepositions poses a very serious problem to machine translation because prepositions are highly ambigous. In theory prepositions can be disambiguated by a filter that excludes already generated representational objects with no selection restriction match between preposition and np, but it takes too long time in practice. A neural network makes the disambiguation in fractions of a second, because it is fast, robust and very powerful.

1 The Problem

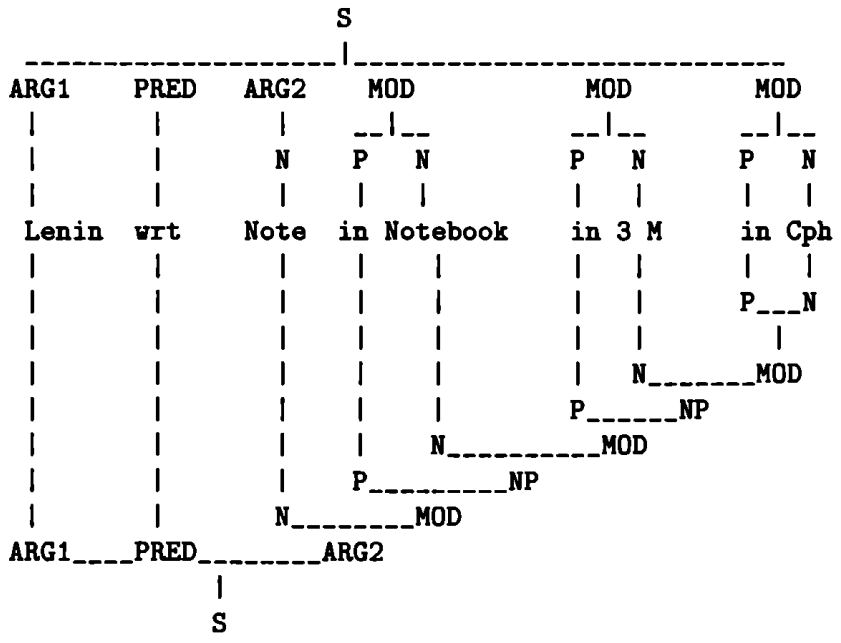
The translation of prepositions poses a very serious problem to machine translation because prepositions are highly ambigious—each of the most 10 frequent prepositions in one of the 9 EUROTRA languages is translated into 10 different prepositions in each of the 8 other languages—and because prepositions always will generate many attachment patterns.

Take the example:

Lenin wrote this note in his notebook in 3 minutes in Copenhagen

There are the flat structure and the deep structure and 6 attachment patterns in between:

Proceedings of NODALIDA 1989, pages 237-249



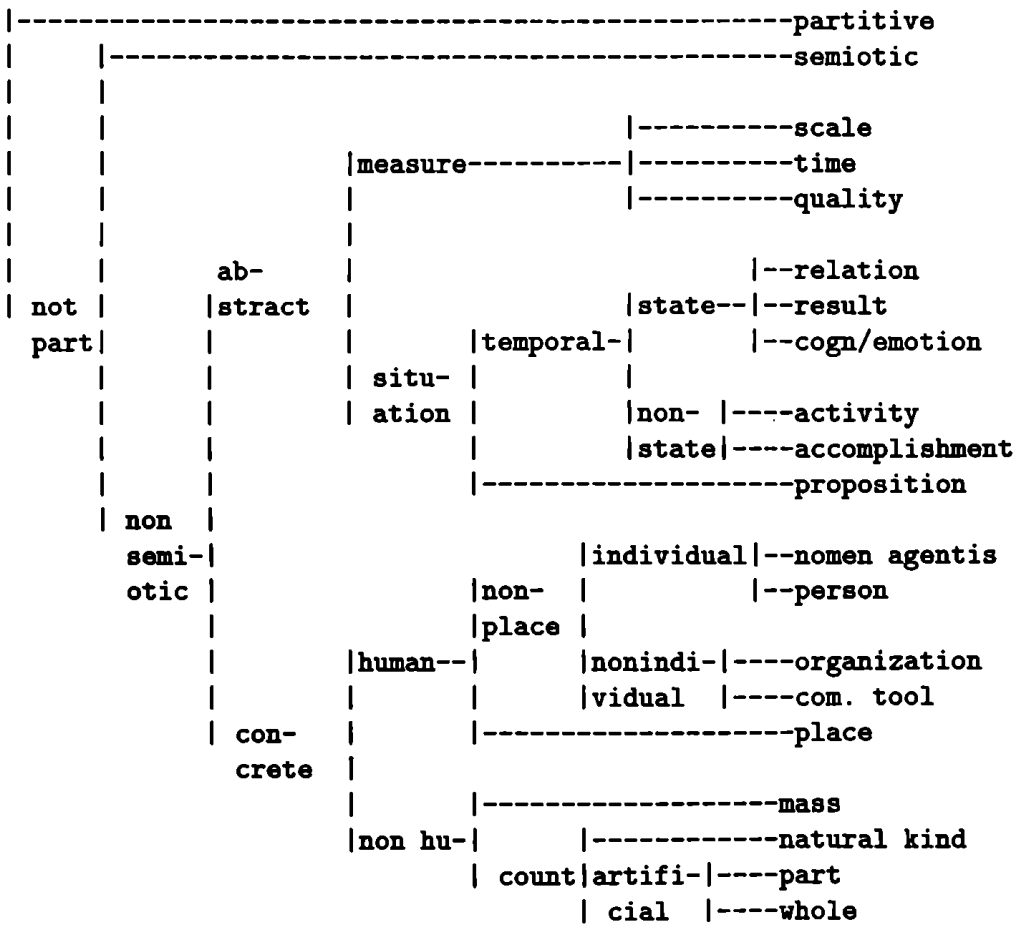
It is obvious in this example that *in* does not mean the same thing in the three clauses: *in his notebook*, *in 3 minutes* and *in Copenhagen*. The first one means DIRECTION, the second one: TIME: HOW LONG, and the third one: PLACE WHERE. In other words, we need (at least) three lexical entries for the word *in*. Other entries are needed for *in 1897*, *in anger*, *in danger*, *in various colours*, with the manings: TIME WHEN, IN A MOOD OF, DURING AN ACTIVITY OF, OF A QUALITY OF. So in Danish there are, in my opinion, at least 12 different meanings of the corresponding preposition *i* (and furthermore all the fixed phrases, e.g. *in all*, *to be in for*).

The distinction between DIRECTION and PLACE WHERE is not made intuitively, but with the so called 'nonsensical conjunction reduction test', which says: if the conjunction of two contexts to the same word does not make sense, there are two readings of the lexical item, one for each context. So it does not make sense to say: *Lenin wrote this note in his notebook and three minutes*, or *Lenin wrote this note in his notebook and Copenhagen*, or *Lenin wrote this note in three minutes and Copenhagen* or *Lenin wrote this note in three minutes and 1897*. But note that contexts of the same type, i.e. contexts of the same reading of a given word form can be coordinated and make sense: *Lenin wrote this note in anger and disappointment.*

If we parse the sentence using a dictionary with 12 different entries for the word *in*, and 8 different attachment patterns, the machine will generate $8 \times 12 \times 12 \times 12 = 13.824$ different representational objects i.e. readings of the sentence. So the problem is how to disambiguate the sentence and make the machine find the correct reading among the 13.824 syntactically possible readings.

2 The Linguistic Solution

In theory, that means from a linguistic point of view, this can be done in the following way: in the dictionary the 12 entries for *in* are differentiated wrt. their **selection restrictions**: in_1 (PLACE WHERE) takes as argument1 nouns of the type CONCRETE, in_2 (TIME WHEN) takes nouns of the type YEAR or DATE, in_3 (MOOD) takes nouns of the type COGNITION OR EMOTION, in_4 (TIME HOW LONG) takes nouns of the type SCALE. That means that all nouns in the dictionary are coded with a semantic type label. I have suggested the following set of semantic types for nouns:



Danish examples: PARTITIVE: *sektor, side, halvdel*, SEMIOTIC: *afsnit, forslag, aftale*, SCALE: *meter, decibel, grad, minut, dag*, TIME: *efterkrigstiden, fremtiden*, QUALITY: *identitet, størrelse, længde*, RELATION: *afhængighed, faktor, position*, RESULT: *produktion_2, undtagelse, investering_2*, COGN/EMOTION: *interesse, frygt, glæde*, ACTIVITY: *databehandling, anvendelse, produktion_1*, ACCOMPLISHMENT: *revolution, investering_1, udforskning*, PROPOSITION nouns: *fordel, mulighed, problem*, NOMINA AGENTIS: *fabrikant af, tilskuer til, herre over*, ORGANIZATION: *hjemmemarked, industri, regering*, COMMUNICATION TOOL: *persondatamat, videobåndoptager, radio*, PLACE: *Europa,*

København, ildlinjen, MASS: *vand, luft, sand*, NATURAL KIND: *blomst, træ, sten*, PART: *kredsløb, svinghjul, taster*, WHOLE: *dataanlæg, elektronik, informationsteknologi.*

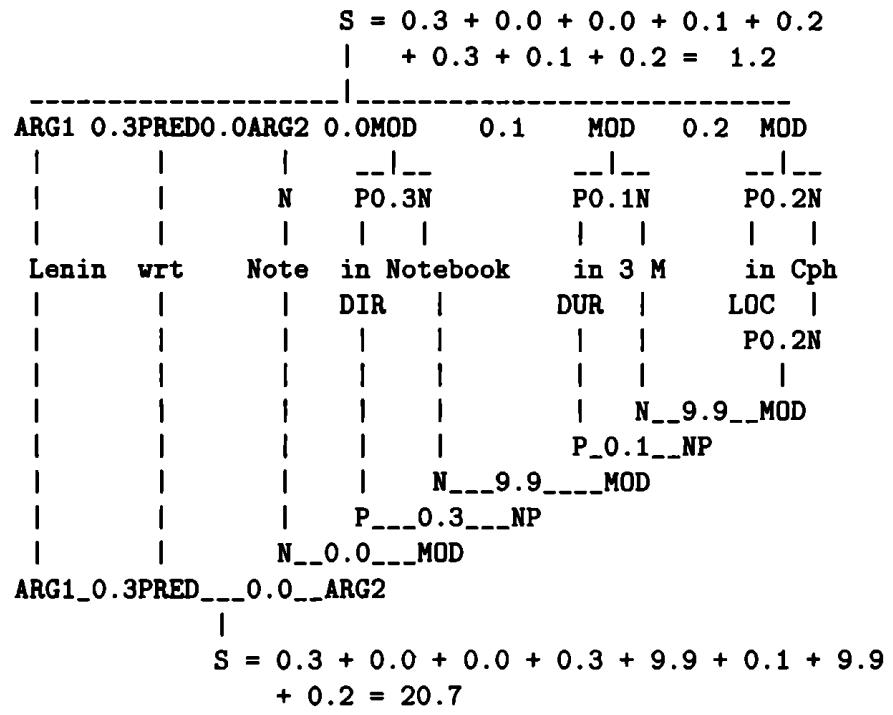
The selection restriction could work as a filter with a 'killer rule', i.e. a rule that would exclude ('kill') all objects with no match between the type which is asked for in the semantic frame specification of the head, in this case the preposition, and the type of the noun that fills the slot. There will be no match in the created object with in_1 (PLACE WHERE) in the clause *in three minutes*, because *minutes* is a noun of the type SCALE, and *in_1* only selects nouns of the type CONCRETE.

This type of rule would exclude most of the not wanted objects among the 13.824 generated representational objects. But the rule is too strong, because it is not uncommon in natural texts to find metaphorical or slightly metaphorical sentences, e.g.: *The situation threatens to become worse.* In this case the selection rule saying that the verb *threaten* only takes nouns of type HUMAN as argument1 will 'kill' all the generated objects so that no analysis or translation will be produced at all.

3 Preference Rules

Instead it is necessary to use a **preference** rule that compares all representational objects generated from the same surface structure, ranks them wrt. *internal semantic fitness*, and selects the fittest. As shown in the first paragraph the simple example *Lenin wrote this note in his notebook in three minutes in Copenhagen* will generate 8 attachment patterns which then can have 12 different readings of each of the three prepositions. What is compared by a preference rule is not two clauses containing the same two or three words, but the sum of the semantic distances between all the pairs in the sentence of 1) a selection restriction bearing head and 2) the corresponding slot filler, added up at the top node.

The concept of semantic distance and semantic fitness can be operationalized in the tree of semantic types. You walk in the tree from the type which is asked for in the selection restriction, step by step, to the type of the slotfiller, counting 1.0 for every step to the left, and 0.1 for every step to the right. The distance from CONCRETE (which is selected by *in_1*, PLACE WHERE) to SCALE (the type of *minutes*) is 1.3, while the distance from SCALE (which is selected by *in_4* TIME HOW LONG) and SCALE is 0.0. Consequently reading *in_4*, TIME HOW LONG is selected in the clause *in three minutes.* Two representational objects, two tree structures representing two whole sentences, can then be compared in the following way:



By such a preference rule the flat structure with the DIRECTION reading, the HOW LONG reading and the WHERE reading, respectively, will be selected—and that is exactly the correct one among the 13.824 possible readings.

But this machinery will only work in theory. The comparison among the objects will be made in pairs, so there will be made $13.824/2 \times 13.825$ comparisons and that will take approximately $6\frac{1}{2}$ hours with a fast machine and a fast program.

4 The Neural Network Design

So in theory it can be done, and the human brain must follow a rule like the one described when it calculates the correct reading in fractions of a second, but it must do it in a smarter way than by comparison in pairs of already genereted objects.

This smarter way must be something like what is called a neural network, which is a strategy for programming the preference rule so that the machine can compute the best solution of the problem in fractions of a second, like the human brain does.

The semantic network is designed in the following way: It consists of three layers, an input layer with 117 neurons, a hidden layer with 65 neurons, and an output layer with 12 neurons. All input neurons are connected with all the hidden layer neurons, and all the hidden layer neurons are connected with all the output layer neurons. That means that there are 7670 connexions between layer 1 and 2, and 792 connexions between layer 2 and 3.

Each of the output neurons represents one of the possible readings of the preposition *in*. The 12 readings of *in* are: ARG1 (deep subject), ARG2 (deep object), LOC (place where), DIR (direction), TIME, DUR (time how long), MEA (measure), STA (state), ACT (activity), EMO (emotion/cognition), QUAL (quality), CLOT (clothes).

The input is a pattern of the syntactic and semantic structure of a sentence containing the word *in*. 4 words to the left and 4 to the right of the preposition are represented in the pattern as syntactic-semantic categories. A given word belongs to one and only one of the following 56 categories, which include the semantic features described in the first paragraph:

```
NOUNS: NONHUMAN              VERBS: AUXILIARY OR MODAL
       PLACE                        INTRANS |  STATE
       HUMAN                        OR PAS- |  PROCESS
       NOMEN AGENTIS                SIVE    |  EVENT
       SEMIOTIC                     TRANSITIVE + noun
       PART                         TRANSITIVE + SENTENCE
       MEASURE OR BARE FORM         TRIVALENT VERB
       TIME                         VERB  with prepositional ob-
       QUALITY OR RELATION           jects and the preposition
       RESULT                        i, til, fra, over, under,
       EMOTION OR COGNITION          for, af, ved
       ACTIVITY
       ACCOMPLISHMENT
       PROPOSITION

PREPOSITIONS: I, PÅ, TIL, FRA, OM, FOR, AF, MED, UDEN, OVER,
              UNDER, MELLEM
PRONOUN                      CONJUNCTION          NUMBER
PUNCTUATION MARK             AND/OR/BUT           TIME ADVERB
THAT                         ARTICLE              PLACE ADVERB
ADJECTIVE                    DIRECTIONAL ADVERB   OTHER ADVERBS
ADJECTIVE + PREPOSITIONAL OBJECT
```

The natural way to represent the 9 word input pattern would be an array with 504 neurons ordered in 9 rows and 56 columns. But that would be a very redundant representation, because only 9 of the 504 neurons would be activated in each sentence.

The input pattern information can be represented by only 117 neurons organized in an array with 9 rows, one for each word in the sentence window, and 13 columns, in which each of the 56 categories is represented by 3 X in accordance with the following coding key (n = noun, v = verb, p = preposition, a = other, o = zero, i = 1):

```
__no_____ni_____vo______vi______po____pi______ao______ai_____
1 nonhum time   aux/mod VPO-i   i     med     pronom  adj    1
2 place  rel    state   VPO-på  på    uden    konj    a-pob  2
3 hum    result process VPOtil  til   over    punkt   tal    3
4 agent  emo    event   VPOov/u fra   under   og      a-tid  4
5 sem    activi t-vb+n  VPOfor  om    mellem  at      a-sted5
6 part   accomp t-bv+s  VPOaf   for   ved     article a-ret  6
7 scale__prop___tri-vb__VPOloc__af____før_____efter___a.adv_7
```

That means that the category INTRANSITIVE VERB OF THE STATE TYPE is represented by *v o 2*. As an example the sentence *Det sker i 1992* ('it happens in 1992') has the representation shown below:

```
SENTENCE: XXX . Det sker  - i - 1992 . XXX XXX
                INPUT PATTERN
                nvpaoi1234567
              -4 .............4-
              -3 ...XX...X....3-
              -2 ...XX.X......2-
              -1 .X..X....X...1-
               0 ..X.X.X......0
              +1 X....X..X....1+
              +2 ...XX...X....2+
              +3 .............3+
              +4 .............4+
                nvpaoi1234567
```

The output is represented by 12 'thermometers' which show how much a given neuron, representing one reading of the preposition *i*, is activated:

```
arg1:     ........
arg2:     ........
 loc:     XXXX....
 dir:     XX......
time:     ........
 dur:     XXXX....
 mea:     ........
 sta:     X.......
 act:     ........
 emo:     ........
qual:     ........
clot:     ........
```

When a pattern of input neurons is activated the neurons 'fire', i e. they activate all the hidden neurons they are connected with, with the weight or strength which is assigned to the specific connexion.

Each of the hidden neurons is now activated by the sum of their input values, which is depending on both the pattern of the firing input neurons and the weight

of the their connexions. The hidden neurons only fire if their activation value exceeds a certain threshold level, and the output neurons are activated in the same way.

5 Rules in Neural Networks

The neural network is 'trained' with examples of input patterns and correct answers. When the training starts all the connexions are randomized, and the output of the network will in the beginning be rather incorrect. Then the correct answer is typed as a second input, and by a process called back propagation all the connexion weights activated by the input sentence are changed. The connexion weights yielding correct output are increased and the connexion weights yielding incorrect output are decreased with a certain rate.

Below I mention some of the 100 Danish input sentences—or rather strings of 9 words, the central word *i* and 4 words to the left and to the right—and the correct answer, i.e. the best reading of the preposition *i* in the context.

20. sikre at hver deltager -i- samme projekt i hele = ARG2
21. deltager i samme projekt -i- hele projektets løbetid til = DUR
22. dominere dette marked og -i- stigende omfang eksportere fra = MEA
23. nu er under overvejelse -i- nogle af de større medlemsstater = LOC
24. til et sådant nyt program -i- stor målestok er kommet = MEA
25. Esprit velkommen til mødet -i- juni 1992 og godkendte = TIME
26. XXX . Det sker -i- 1992 . XXX XXX = TIME
27. som anvendes i operationer -i- mange versioner og varianter = MEA
28. og afprøvning af VLSI/systemer -i- cilisium eller andre halvledere = LOC
29. beslutning end en afgørelse -i- rådet = LOC
30. fuldt kan støtte brugeren i kommunikationsprocessen, og som = ACT
31. ; de vil resultere i nye produkter, processer = ARG2
32. anvendelse foregår meget Langsommere i Europa end i Japan = LOC
33. af alle varer fremstillet i fællesskabet er i små = ARG2

When the connexion strengths have been adjusted a number of times with a number of input sentences the pattern of the connexion strengths will represent a rule wich will yield the correct output to each of the input patterns in the training set.

It is essential that the input sentences are authentic and not grammar book sentences, because all regularities in the input material, even the number of words from the word *i* to the punctuation mark, will be made into a rule by the network.

It is essential too that the number of input sentences is so high that all non-linguistic regularities of any kind are excluded. 100 input sentences are certainly not enough to make sure that all nonimportant word types have been placed in all 8 positions in the input picture.

I am not sure that a window of 9 words is enough, but in the first 100 authentic 9 word input sentences the rule triggering word has been present.

The training can be seen from two screen pictures, the first one showing input pattern, aswer and output of run no. 2 of sentence no. 26. The output is not even slightly in the right direction.

```
------------------------------------------------------------
fact no. 26                                        run no. 2
SENTENCE: XXX . Det sker  - i - 1992 . XXX XXX
   INPUT PATTERN              ANSWER        OUTPUT
   nvpaoi1234567      arg1:   ........      ........
-4 ............4-     arg2:   ........      ........
-3 ...XX...X....3-      loc:   ........      ........
-2 ...XX.X......2-      dir:   ........      XX......
-1 .X..X....X...1-     time:   XXXXXXXX      ........
 0 ..X.X.X......0       dur:   ........      ........
+1 X....X..X....1+      mea:   ........      ........
+2 ...XX...X....2+      sta:   ........      ........
+3 ............3+      act:   ........      X.......
+4 ............4+      emo:   ........      ........
   nvpaoi1234567       qual:   ........      ........
                       clot:   ........      ........
------------------------------------------------------------
```

But in run nr. 15 the network has 'learned' a rule completely, and gives the correct output to all the training sentences.

```
------------------------------------------------------------
fact no. 26                                       run no.15
SENTENCE: XXX . Det sker  - i - 1992 . XXX XXX
   INPUT PATTERN              ANSWER        OUTPUT
   nvpaoi1234567      arg1:   ........      ........
-4 ............4-     arg2:   ........      ........
-3 ...XX...X....3-      loc:   ........      ........
-2 ...XX.X......2-      dir:   ........      ........
-1 .X..X....X...1-     time:   XXXXXXXX      XXXXXXXX
 0 ..X.X.X......0       dur:   ........      ........
+1 X....X..X....1+      mea:   ........      ........
+2 ...XX...X....2+      sta:   ........      ........
+3 ............3+      act:   ........      ........
+4 ............4+      emo:   ........      ........
   nvpaoi1234567       qual:   ........      ........
                       clot:   ........      ........
------------------------------------------------------------
```

It is interesting that the established rule will give the correct answer to new sentences too, i.e. sentences which have never been given as input pattern before. In a way the network has 'learned' a linguistic rule inductively although it has not been formulated explicitly. It can be seen in the following three examples.

```
------------------------------------------------------------
new fact
SENTENCE: Mødet i Strassbourg varede - i - 3 uger
   INPUT PATTERN                 ANSWER        OUTPUT
   nvpaoi1234567       arg1:     ........      ........
-4 X....X....X..4-     arg2:     ........      ........
-3 ..X.X.X......3-      loc:     ........      ........
-2 X...X..X.....2-      dir:     ........      XX......
-1 .X..X...X....1-     time:     ........      ........
 0 ..X.X.X......0       dur:     ........      XXXXXXXX
+1 ...X.X..X....1+      mea:     ........      ........
+2 X...X.......X2+      sta:     ........      ........
+3 .............3+      act:     ........      ........
+4 .............4+      emo:     ........      ........
   nvpaoi1234567       qual:     ........      ........
                       clot:     ........      ........
------------------------------------------------------------


------------------------------------------------------------
new fact
SENTENCE: fordi den fortsatte deltagelse - i - forhandlingerne
         med de implicerede
   INPUT PATTERN                 ANSWER        OUTPUT
   nvpaoi1234567       arg1:     ........      ........
-4 ...XX..X..X..4-     arg2:     ........      XXXXXXX.
-3 ...XX.X......3-      loc:     ........      ........
-2 ...X.XX......2-      dir:     ........      ........
-1 X....X....X..1-     time:     ........      ........
 0 ..X.X.X......0       dur:     ........      ........
+1 X....X....X..1+      mea:     ........      ........
+2 .....XX.....X2+      sta:     ........      ........
+3 ...XX.X......3+      act:     ........      ........
+4 ...X.XX......4+      mea:     ........      ........
   nvpaoi1234567       qual:     ........      ........
                       clot:     ........      ........

------------------------------------------------------------
```

```
------------------------------------------------------------
new fact
SENTENCE: for en tredje rekvirent i samarbejde med en virksom
          hed
   INPUT PATTERN                ANSWER          OUTPUT
   nvpaoi1234567        arg1:   ........        ........
-4 ..X.X......X.4-      arg2:   ........        XX......
-3 ...XX......X.3-       loc:   ........        X.......
-2 ...X.XX......2-       dir:   ........        X.......
-1 X...X....X...1-      time:   ........        ........
 0 ..X.X.X......0        dur:   ........        ........
+1 X....X....X..1+       mea:   ........        ........
+2 ..X..XX......2+       sta:   ........        ........
+3 ...XX......X.3+       act:   ........        XXXXXXX.
+4 X...X...X....4+       emo:   ........        ........
   nvpaoi1234567        qual:   ........        ........
                        clot:   ........        ........
------------------------------------------------------------
```

I have not yet—efter 100 input sentences—statistics about how many percent of correct 'guesses' the network will make about new sentences, but it is already clear that it is possible to make a network which can solve the problem of disambiguation of prepositions without the enormous overgeneration which is made by filter rules in serial programming.

It should according to the theories be possible to train the same network to make the disambiguation of all the prepositions (or all the most frequent and ambiguous prepositions). The network I have described is in fact designed to compute 15 different prepositions. But I have not yet trained it with other prepositions than *i*.

6 The Power of Neural Networks

I imagine that the neural network in the translation process will be placed before the parser. The network is fed with the lexical words of the input sentence, and the relevant information about the semantic type of each word taken from the dictionary. All the prepositions in the sentence are then disambiguated by the network and the reading number asigned to them before they are parsed by the grammar parser. The product of the network would in the example from the beginning of this article be:

> *Lenin wrote this note in(DIR) his notebook in(DUR) 3 minutes in(LOC) Copenhagen in(TIME) 1897 in(EMO) anger.*

The enormous disambiguation power of the neural network results from three factors: the parallel distribution, which makes it fast, the nonlocal representation

of the rule, which makes it robust, and the statistical analysis, which makes it powerful.

The machine does not in fact compute the rule in parallel, but in a serial machine the program simulates the parallel processing, and that is enough to compute the disambiguation of a preposition in fractions of a second. 8462 calculations do not take more than a fraction of a second.

The rules which are used for disambiguation of the preposition *i*, one of which could be that *i* followed by a noun of the semantic type PLACE will normally be a *i(LOC)*, are not located in some of the connexions, but in the whole pattern of connexions both from input layer to hidden layer, and from hidden layer to output layer. So irregularities in the input, metaphors or syntactic errors, will not totally disable the rule, but only make minor changes in the output. The network will always find the 'best' solution, i.e. recognize the reading with most semantic fitness regardless how good or bad it is—exactly as we do even when we read the famous sentence: *Colorless green ideas sleep furiously.*

The nonlocal representation offers a solution of the problem of the so called hermeneutic circle, the problem that the whole can not be understood before the parts are understood, and the parts can not be understood before the whole is understood. The meaning of the sentence consists of, but is at the same time more than the sum of the senses of the words.

With nonlocal representations the meaning of the whole is represented, not as the sum of the meaning of the parts, but as a pattern or 'meaning' of something which is subsymbolic, subsignificant or with no meaning at all, but with a differentiating function, viz. the neurons of the hidden layer. So the network computes or recognizes the meaning of the whole by computing, not the sum of the parts, but the pattern of the subsymbolic parts (the hidden layer neurons) of the symbolic parts (the words) of the sentence.

That is exactly the function of letters or phonemes, which have no meaning but only a differentiation function, and nevertheless make it possible to transmit word senses and sentence meaning from sender to receiver in the communication process between humans.

But most important, the neural network will utilize information which can not be used in normal grammar rules, viz. probabilistic infomation. It is a linguistic rule that only *in(DUR)* will be followed by a noun of the type SCALE: *in 3 minutes.* Let us assume that it is a statistical rule that *in(DUR)* is followed by a cardinal number 1.000 times more often than *in(LOC)* is. It is not possible to formulate this regularity as a linguistic rule, not even as a preference rule, because of the possibility of the sentence: *she worked in two rooms.* The semantic network will utilize the probabilistic information but not make errors in this crucial example, because the pattern of connexion weights has learned the rule for the combination of cardinal numbers and measure nouns, not for cardinal numbers only.

References

Rumelhart, David E., James L. McClelland and the PDP sesearch Group. 1986. *Parallel distributed processing. Explorations in the Microstrucure of Cognition.* Vols. 1–3. MIT Press, Cambridge, Mass.

EUROTRA-DK
Njalsgade 80
DK-2300 Copnehagen S
Denmark